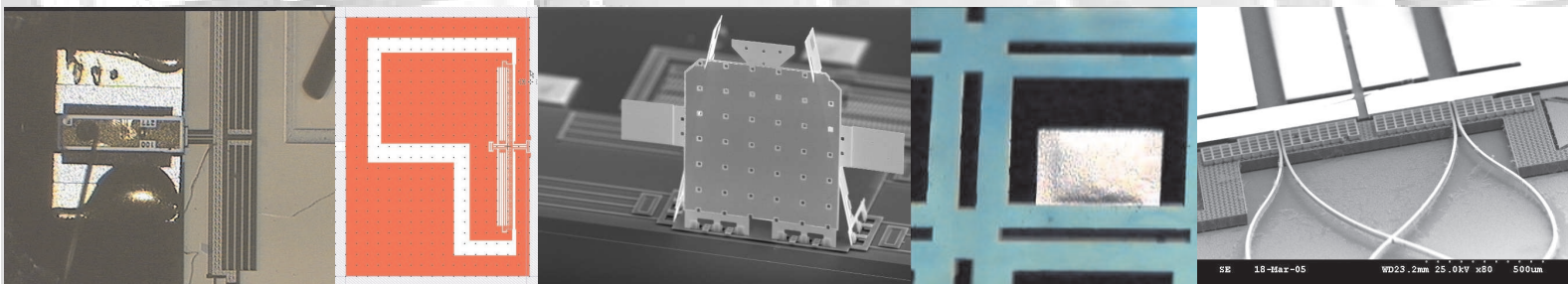


# A (not so) short introduction to MEMS

Franck CHOLLET, Haobing LIU



— 5.4 —

[memscyclopedia.org](http://memscyclopedia.org)

The original source for the document you are viewing has been written with L<sup>A</sup>T<sub>E</sub>X. The diagrams in the book were mostly created using Inkscape and CorelDraw<sup>©</sup> sometimes incorporating graphs produced with MATLAB<sup>©</sup>. The photographs were processed (mostly contrast enhancement) when required with GIMP.

The title of the book is an homage to the famous introduction to L<sup>A</sup>T<sub>E</sub>X by Tobias Oetiker.

All diagrams and most photographs are original to this book. Additional photographs are licensed from a variety of sources as indicated in the caption below the figure. These additional photographs can not be reused outside of this book without asking the original copyright owners.

For the full copyright of the original material (photographs, diagrams, graphs, code and text) see the following page.

LIBRARY CLASSIFICATION (DEWEY): 621.3

ISBN: 978-2-9542015-0-4

ISBN 978-2-9542015-0-4



Please note that this work is published under a :



Attribution-NonCommercial 3.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:



Attribution.

Please attribute this work using: “A (not so) short Introduction to Micro Electromechanical Systems”, F. Chollet, HB. Liu, version 5.4, 2018, <<http://memscyclopedia.org/introMEMS.html>>.



Noncommercial.

You may not use this work for commercial purposes.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holders, who can be contacted at <http://memscyclopedia.org/contactus.html>.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code  
(<http://creativecommons.org/licenses/by-nc/3.0/legalcode>).



# Contents

<b>Contents</b>	<b>5</b>
<b>1 Why MEMS?</b>	<b>9</b>
1.1 What is MEMS and comparison with microelectronics . . . . .	9
1.2 Why MEMS technology . . . . .	10
1.2.1 Advantages offered . . . . .	10
1.2.2 Diverse products and markets . . . . .	12
1.2.3 Economy of MEMS manufacturing and applications . . . . .	14
1.3 Major drivers for MEMS technology . . . . .	16
1.4 Mutual benefits between MEMS and microelectronics . . . . .	17
<b>2 Introduction to MEMS modeling</b>	<b>19</b>
2.1 Physical scaling laws . . . . .	19
2.2 The principles of design and reliability . . . . .	22
2.3 MEMS design tools . . . . .	24
2.4 Lumped elements modeling . . . . .	26
2.4.1 Block modeling . . . . .	29
2.4.2 Open and closed-loop systems . . . . .	30
2.4.3 Linear system and Laplace's transform . . . . .	33
2.4.4 Analogies and circuit modeling . . . . .	36
2.5 Dynamic analysis . . . . .	39
2.5.1 Time domain analysis with Laplace's transform . . . . .	40
2.5.2 Frequency domain analysis with Fourier's transform . . . . .	41
2.5.3 First-order model . . . . .	44
2.5.3.1 Step-response . . . . .	45
2.5.3.2 Frequency response . . . . .	46
2.5.4 Second-order model . . . . .	47
2.5.4.1 Step-response . . . . .	47
2.5.4.2 Frequency response . . . . .	49
2.5.4.3 Quality factor . . . . .	52
2.5.5 Effect of frequency on system response . . . . .	52
2.5.5.1 Gain distortion . . . . .	52
2.5.5.2 Phase distortion . . . . .	54
2.6 Advanced sub-systems modeling . . . . .	57

2.6.1	Multi-domain circuit elements . . . . .	57
2.6.2	Non-linear sub-system dynamics . . . . .	58
Problems	. . . . .	61
<b>3</b>	<b>How MEMS are made</b>	<b>63</b>
3.1	Overview of MEMS fabrication process . . . . .	63
3.2	MEMS materials . . . . .	69
3.2.1	Crystalline, polycrystalline and amorphous materials . . . . .	69
3.2.2	Materials properties . . . . .	73
3.3	Vacuum technology . . . . .	79
3.3.1	Vacuum actuators . . . . .	83
3.3.1.1	Rotary vane pump . . . . .	85
3.3.1.2	Roots pump . . . . .	85
3.3.1.3	Scroll pump . . . . .	86
3.3.1.4	Diffusion pump . . . . .	87
3.3.1.5	Turbo pump . . . . .	87
3.3.1.6	Cryo pump . . . . .	89
3.3.2	Vacuum sensors . . . . .	90
3.3.2.1	Membrane gauge . . . . .	92
3.3.2.2	Pirani gauge . . . . .	93
3.3.2.3	Ionization gauge . . . . .	94
3.4	Bulk micromachining, wet and dry etching . . . . .	96
3.4.1	Isotropic and anisotropic wet etching . . . . .	96
3.4.2	Dry etching . . . . .	100
3.4.3	Wafer bonding . . . . .	104
3.5	Surface micromachining and thin-films . . . . .	106
3.5.1	Thin-film fabrication . . . . .	107
3.5.1.1	Oxidation . . . . .	109
3.5.1.2	Doping by diffusion and ion implantation . . . . .	111
3.5.1.3	Spin-coating . . . . .	114
3.5.1.4	Physical Vapor Deposition (PVD) techniques . . . . .	116
3.5.1.5	Chemical Vapor Deposition (CVD) techniques . . . . .	118
3.5.1.6	Epitaxy . . . . .	120
3.5.2	Design limitation . . . . .	121
3.5.3	Microstructure release . . . . .	123
3.6	DRIE micromachining . . . . .	125
3.7	Other microfabrication techniques . . . . .	129
3.7.1	Micro-molding and LIGA . . . . .	129
3.7.2	Polymer MEMS . . . . .	130
3.8	Characterization . . . . .	131
3.8.1	Light Microscope . . . . .	131
3.8.2	SEM (Scanning Electron Microscope) . . . . .	142
3.8.3	Contact probe profilometry . . . . .	146
Problems	. . . . .	149

Solutions . . . . .	152
<b>4 MEMS technology</b>	<b>153</b>
4.1 MEMS system partitioning . . . . .	153
4.2 Fabrication tolerance and design rules . . . . .	155
4.2.1 Fabrication tolerance . . . . .	155
4.2.2 Design rules . . . . .	156
4.3 Passive structures . . . . .	160
4.3.1 Mechanical structures . . . . .	160
4.3.2 Distributed mechanical structures . . . . .	165
4.3.3 Fluidic structures . . . . .	168
4.3.3.1 Static properties . . . . .	170
4.3.3.2 Dynamic properties . . . . .	178
4.4 Sensor technology . . . . .	193
4.4.1 Piezoresistive sensing . . . . .	193
4.4.2 Capacitive sensing . . . . .	196
4.4.3 Other sensing mechanism . . . . .	200
4.5 Actuator technology . . . . .	201
4.5.1 Magnetic actuator . . . . .	202
4.5.2 Electrostatic actuator . . . . .	203
4.5.3 Piezoelectric actuator . . . . .	208
4.5.4 Thermal actuator . . . . .	212
Problems . . . . .	219
Solutions . . . . .	223
<b>5 MEMS packaging, assembly and test</b>	<b>229</b>
5.1 Assembly . . . . .	231
5.2 Packaging . . . . .	232
5.2.1 Encapsulation . . . . .	235
5.2.2 Hermetic encapsulation . . . . .	238
5.2.3 Electrical feedthrough . . . . .	247
5.3 Testing and calibration . . . . .	249
5.3.1 Testing . . . . .	251
5.3.2 Calibration . . . . .	252
5.3.3 Compensation . . . . .	254
Problems . . . . .	262
<b>6 Challenges, trends, and conclusions</b>	<b>263</b>
6.1 MEMS current challenges . . . . .	263
6.2 Future trends in MEMS . . . . .	264
6.3 Conclusion . . . . .	264

<b>A Readings and References</b>	<b>267</b>
A.1 Conferences . . . . .	267
A.2 Online resources and journals . . . . .	268
A.3 Other MEMS resources . . . . .	269
<b>B Causality in linear systems</b>	<b>271</b>
<b>C Resonator and quality factor</b>	<b>273</b>
<b>D Laplace's transform</b>	<b>275</b>
<b>E Complex numbers</b>	<b>277</b>
<b>F Fraunhofer diffraction</b>	<b>281</b>
F.1 Far-field diffraction . . . . .	281
F.2 Bessel function . . . . .	283
<b>G OCTAVE code</b>	<b>285</b>
G.1 Bode diagram . . . . .	285
<b>Bibliography</b>	<b>289</b>
<b>Index</b>	<b>293</b>



# Chapter 1

## Why MEMS?

### 1.1 What is MEMS and comparison with microelectronics

Micro Electro Mechanical Systems or MEMS is a term coined around 1989 by Prof. R. Howe [1] and others to describe an emerging research field, where mechanical elements, like cantilevers or membranes, had been manufactured at a scale more akin to microelectronics circuit than to lathe machining. But MEMS is not the only term used to describe this field and from its multicultural origin it is also known as Micromachines, a term often used in Japan, or more broadly as Microsystems Technology (MST), in Europe.

However, if the etymology of the word is more or less well known, the dictionaries are still mum about an exact definition. Actually, what could link inkjet printer head, video projector DLP system, disposable bio-analysis chip and airbag crash sensor - yes, they are all MEMS, but what is MEMS?

It appears that these devices share the presence of features below  $100\ \mu\text{m}$  that are not machined using standard machining but using other techniques globally called micro-fabrication technology.

“Micro-fabrication makes the MEMS.”

Of course, this simple definition would also include microelectronics – and some would do it in the broader microsystem term – but there is a characteristic that electronic circuits do not share with MEMS. While electronic circuits are inherently solid and compact structures, MEMS have holes, cavity, channels, cantilevers, membranes, etc, and, in some way, imitate ‘mechanical’ parts.

This has a direct impact on their manufacturing process. Actually, even when MEMS are based on silicon, microelectronics process needs to be adapted to cater for thicker layer deposition, deeper etching and to introduce special steps to free the mechanical structures. Then, many more MEMS are not based on silicon and can be manufactured in polymer, in glass, in quartz or even in metals...

Thus, if similarities between MEMS and microelectronics exist, they now clearly are two distinct fields. Actually, MEMS needs a completely different set of mind,

where next to electronics, mechanical and material knowledge plays a fundamental role.

## 1.2 Why MEMS technology

### 1.2.1 Advantages offered

The development of a MEMS component has a cost that should not be underestimated, but the technology has the possibility to bring unique benefits. The reasons that prompt the use of MEMS technology can be classified broadly in three classes:

**miniaturization of existing devices** For example the production of silicon based gyroscope which reduced existing devices weighting several kg and with a volume of  $1000\text{cm}^3$  to a chip of a few grams contained in a  $0.5\text{cm}^3$  package.

**using physical principles that do not work at larger scale** A typical example is given by the biochips where electric field are use to pump the reactant around the chip. This so called electro-osmotic effect based on the existence of a drag force in the fluid works only in channels with dimension of a fraction of one mm, that is, at micro-scale.

**developing tools for operation in the micro-world** In 1986 H. Rohrer and G. Binnig at IBM were awarded the Nobel price in physics for their work on scanning tunneling microscope. This work heralded the development of a new class of microscopes (atomic force microscope, scanning near-field optical microscope...) that shares the presence of micromachined sharp micro-tips with radius below 50 nm. This micro-tool was used to position atoms in complex arrangement, writing Chinese character or helping verify some prediction of quantum mechanics. Another example of this class of MEMS devices at a slightly larger scale would be the development of micro-grippers to handle cells for analysis.

By far miniaturization is often the most important driver behind MEMS development. The common perception is that miniaturization reduces cost, by decreasing material consumption and allowing batch fabrication, but an important collateral benefit is also in the increase of applicability. Actually, reduced mass and size allow placing the MEMS in places where a traditional system won't have been able to fit. Finally, these two effects concur to increase the total market of the miniaturized device compared to its costlier and bulkier ancestor. A typical example is the case of the accelerometer shown in Figure 1.1. From humble debut as crash sensor, it was used in high added value product for image stabilization, then in game controller integrated inside the latest handphonedes, until the high volume and low price allowed it to enter the toys market.

However often miniaturization alone cannot justify the development of new MEMS. After all if the bulky component is small enough, reliable enough, and

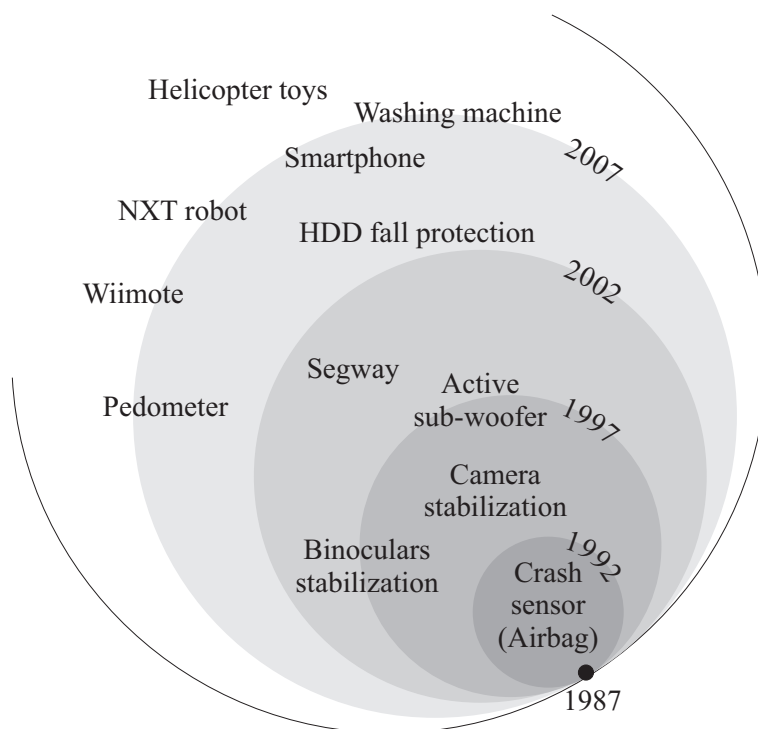


Figure 1.1: Increased use of accelerometers in consumer products.

particularly cheap then there is probably no reason to miniaturize it. Micro-fabrication process cost cannot usually compete with metal sheet punching or other conventional mass production methods.

But MEMS technology allows something different, at the same time you make the component smaller you can make it smarter. The airbag crash sensor gives us a good example of the added value that can be brought by developing a MEMS device. Some non-MEMS crash sensors are based on a metal ball retained by a rolling spring or a magnetic field. The ball moves in response to a rapid car deceleration and shorts two contacts inside the sensor. A simple and cheap method, but the ball can be blocked or contact may have been contaminated and when you start your engine, there is no easy way to tell if the sensor will work or not. MEMS devices can have a built-in self-test feature, where a micro-actuator will simulate the effect of deceleration and allow checking the integrity of the system every time you start the engine.

Another advantage that MEMS can bring relates with the system integration. Instead of having a series of external components (sensor, inductor...) connected by wire or soldered to a printed circuit board, the MEMS on silicon can be integrated directly with the electronics. Whether it is on the same chip or in the same package it results in increased reliability and decreased assembly cost, opening new application opportunities.

MEMS technology not only makes things smaller but often makes them better.

## 1.2.2 Diverse products and markets

The previous difficulty we had to define MEMS stems from the vast number of products that fall under the MEMS umbrella since the first commercial products appeared in 1973. The timescale in Figure 1.2 reveals the emergence of the main successful MEMS products since that early time - and if it seems to slow lately, it is only because we choose to show only those products that are here to stay and have generally changed the market landscape, which takes a few years to prove.

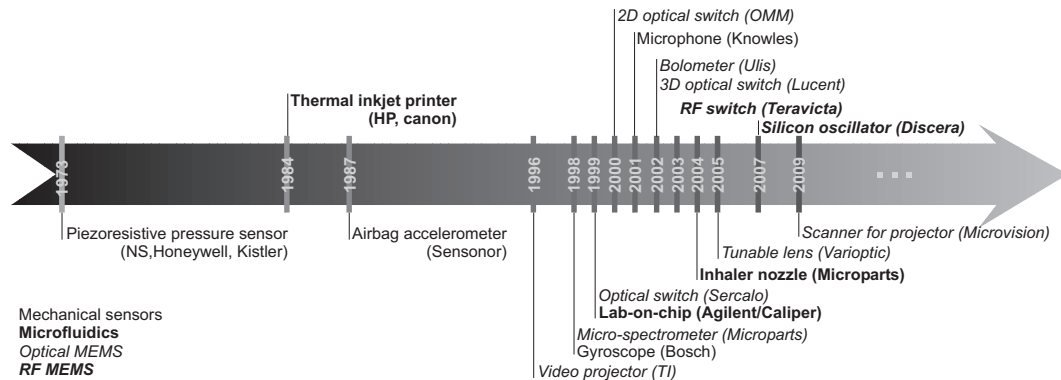


Figure 1.2: Timescale of MEMS products emergence.

The MEMS component currently on the market can be broadly divided in six categories (Table 1.1), where next to the well-known pressure and inertia sensors produced by different manufacturer like Motorola, Analog Devices, Sensoror or Delphi we have many other products. The micro-fluidic application are best known for the inkjet printer head popularized by Hewlett Packard, but they also include the bioMEMS market with micro analysis system like the capillary electrophoresis system from Agilent or the DNA chips.

Optical MEMS (also called MOEMS) includes the component for the fiber optic telecommunication like the switch based on a moving mirror produced by Sercalo. They also include the optical switch matrices that have not yet emerged from the telecommunication industry bubble at the beginning of the 21st century. This component consists of 100s of micro-mirror that can redirect the light from one input fiber to one output fiber, when the fibers are arranged either along a line (proposed by the now defunct Optical Micro Machines) or in a 2D configuration (Lambda router from Lucent). Moreover MOEMS deals with the now rather successful optical projection system that is competing with the LCD projector. The MEMS products are based either on an array of torsional micro-mirror in the Texas Instruments Digital Light Processor (DLP) system or on an array of controllable grating as in the Grating Light Valve (GLV) from Silicon Light Machines. RF MEMS is also emerging as a viable MEMS market. Next to passive components like high-Q inductors produced on the IC surface to replace the hybridized component as proposed by MEMSCAP we find RF switches, silicon oscillators (SiTime) and soon micromechanical filters.

But the list does not end here and we can find micromachined relays (MMR) produced for example by Omron, HDD read/write head and actuator or even toys, like the autonomous micro-robot EMRoS produced by Epson.

<b>Product type</b>	<b>Examples</b>
Pressure sensor	Manifold pressure (MAP), tire pressure, blood pressure..
Inertia sensor	Accelerometer, gyroscope, crash sensor...
Microfluidics / bioMEMS	Inkjet printer nozzle, micro-bio-analysis systems, DNA chips...
Optical MEMS / MOEMS	Micro-mirror array for projection (DLP), micro-grating array for projection (GLV), optical fiber switch, adaptive optics...
RF MEMS	High Q-inductor, switches, antenna, filter..
Others	Relays, microphone, data storage, toys...

Table 1.1: MEMS products example

In 2010 these products represented a market of a bit less than 9B\$, increasing by almost 200% since 2002, with roughly 16% each for the traditional inkjet printer nozzle and pressure sensor market, 33% in inertial sensors (accelerometers and gyroscope), 12% each in projection display and microfluidics and the rest split between RF MEMS, microphones, oscillators[2]... Of course the MEMS market overall value is still small compared to the 300B\$ IC industry - but there are two aspects that make it very interesting for investors:

- it is expected to grow annually at a 2 digit rate for the foreseeable future, much higher than any projection for the mature IC industry market;
- MEMS chips have a large leveraging effect, and in the average a MEMS based systems will have 8 times more value than the MEMS chip price (e.g., a DLP projector is about 10 times the price of a MEMS DLP chip).

It should be noted that this last point has created large discrepancies between market studies, whether they report the market for components alone or for devices. The number we cited above are in the average of other studies and represent the market for the MEMS part alone (actually a somewhat fairer comparison with electronics industry – where device price is considered – would put the value of the MEMS market to more than 70B\$).

### 1.2.3 Economy of MEMS manufacturing and applications

However large the number of opportunities is, it should not make companies believe that they can invest in any of these fields randomly. For example, although the RF MEMS market is growing fueled by the appetite for smaller wireless communication devices, it seems to grow mostly through internal growth. Actually the IC foundries are developing their own technology for producing, for example, high-Q inductors, and it seems that an external provider will have a very limited chance to penetrate the market.

Thus, market opportunities should be analyzed in detail to eliminate the false perception of a large market, taking into consideration the targeted customer inertia to change – and the possibility that the targeted customer develops by himself a MEMS based solution! In that aspect, sensors seems more accessible being simple enough to allow full development within small business unit and having a large base of customers - in the other hand, an optical switch matrix is riskier because its value is null without the system that is built by a limited number of companies, which, most probably, also have the capabilities to develop in-house the MEMS component...

Some MEMS products already achieve high volume and benefit enormously from the batch fabrication techniques. For example more than 100 millions MEMS accelerometers are sold every year in the world - and with newer use coming, this number is still growing fast. But large numbers in an open market invariably means also fierce competition and ultimately reduced prices. Long are gone the days where a MEMS accelerometer could be sold 10\$ a piece - it is now less than 2\$ and still dropping. Currently, the next target is a 3-axis accelerometer in a single package for about 4\$, so that it can really enter the toys industry. And don't expect that you will be able to ramp -up production and decrease prices easily : many of the initial producers of MEMS accelerometers(e.g. Novasensor) have not been able to survive when the price went south of 5\$ as their design could not be adapted to lower production cost. New companies overtook them (e.g. ST Microelectronics), with design aimed from the start at reaching the 1\$ mark...

Of course, there are a few exceptions to this cost rule. Actually, if the number of unit sold is also very large, the situation with the inkjet printer nozzle is very different. Canon and Hewlett Packard developed a completely new product, the inkjet printer, which was better than earlier dot matrix printer, and created a new captive market for its MEMS based system. This has allowed HP to repeatedly top the list of MEMS manufacturer with sales in excess of 600M\$. This enviable success will unfortunately be hard to emulate – but it will be done again!

But these cases should not hide the fact that MEMS markets are often niche markets. Few product will reach the million unit/year mark and in 2006 among the more than 300 companies producing MEMS only 18 had sales above 100m\$/year. Thus great care should be taken in balancing the research and development effort, because the difficulty of developing new MEMS from scratch can be daunting and the return low. Actually current customers expect very high reliability, with

<10 ppm failed parts for consumer products, and even <1 ppm for automotive applications. As such, it is not surprising that a ‘normal’ component development time for automotive applications, as acknowledged by Sensoror, would be 2-3 years – and 2-3 years more if substantial process development is required... And it may be worse. Although Texas Instruments is now reaping the fruit of its Digital Light Processor selling between 1996 and 2004 more than 4 millions chips for a value now exceeding 200m\$/year, the development of the technology by L. Hornbeck took more than 10 years [3]. Few startup companies will ever have this opportunity – and don’t be over-optimistic: when a prototype is done only 10-20% of the job is done.

Actually it is not clear for a company what is the best approach for entering the MEMS business, and we observe a large variety of business model with no clear winner. For many years in microelectronics industry the abundance of independent foundries and packaging companies has made fabless approach a viable business model. However it is an approach favored by only a handful of MEMS companies and as it seems now, for good reasons.

A good insight of the polymorphic MEMS business can be gained by studying the company MemsTech, now a holding listed on the Kuala Lumpur Mesdaq (Malaysia) and having office in Detroit, Kuala Lumpur and Singapore.

Singapore is actually where everything started in the mid-90’s for MemsTech with the desire from an international company (EG&G) to enter the MEMS sensor market. They found a suitable partner in Singapore at the Institute of Microelectronics (IME), a research institute with vast experience in IC technology.

This type of cooperation has been a frequent business model for MNC willing to enter MEMS market, by starting with ex-house R&D contract development of a component. EG&G and IME designed an accelerometer, patenting along the way new fabrication process and developing a cheap plastic packaging process. Finally the R&D went well enough and the complete clean room used for the development was spun-off and used for the production of the accelerometer.

Here, we have another typical startup model, where IP developed in research institute and university ends up building a company. This approach is very typical of MEMS development, with a majority of the existing MEMS companies having been spun-off from a public research institute or a university.

A few years down the road the fab continuously produced accelerometer and changed hands to another MNC before being bought back in 2001 by its management. During that period MemsTech was nothing else but a component manufacturer providing off-the-shelf accelerometer, just like what Motorola, Texas Instrument and others are doing.

But after the buyout, MemsTech needed to diversify its business and started proposing fabrication services. It then split in two entities: the fab, now called Sensfab, and the packaging and testing unit, Senzpak. Three years later, the company had increased its ‘off-the-shelf’ product offering, proposing accelerometer, pressure sensor, microphones and one IR camera developed in cooperation with local and

overseas university.

This is again a typical behavior of small MEMS companies where growth is fueled by cooperation with external research institutions. Still at the same time MemsTech proposes wafer fabrication, packaging and testing services to external companies. This model where products and services are mixed is another typical MEMS business model, also followed by Silicon Microstructures in the USA, Colybris in Switzerland, MEMSCAP in France and some other. Finally, in June 2004 MemsTech went public on the Mesdaq market in Kuala Lumpur.

The main reason why the company could survive its entire series of avatars, is most probably because it had never overgrown its market and had the wisdom to remain a small company, with staff around 100 persons. Now, with a good product portfolio and a solid base of investor it is probably time for expansion.

### 1.3 Major drivers for MEMS technology

From the heyday of MEMS research at the end of the 1960s, started by the discovery of silicon large piezoresistive effect by C. Smith[4] and the demonstration of anisotropic etching of silicon by J. Price[5] that paved the way to the first pressure sensor, one main driver for MEMS development has been the automotive industry. It is really amazing to see how many MEMS sensor a modern car can use! From the first oil pressure sensors, car manufacturer quickly added manifold and tire pressure sensors, then crash sensors, one, then two and now up to five accelerometers. Recently the gyroscopes made their apparition for anti-skidding system and also for navigation unit - the list seems without end.

Miniaturized pressure sensors were also quick to find their ways in medical equipment for blood pressure test. Since then biomedical application have drained a lot of attention from MEMS developer, and DNA chip or micro-analysis system are the latest successes in the list. Because you usually sell medical equipment to doctors and not to patients, the biomedical market has many features making it perfect for MEMS: a niche market with large added value.

Actually cheap and small MEMS sensors have many applications. Digital cameras have been starting using accelerometer to stabilize image, or to automatically find image orientation. Accelerometers are now being used in contactless game controller or mouse.

These two later products are just a small part of the MEMS-based system that the computer industry is using to interface the arid beauty of digits with our human senses. The inkjet printer, DLP based projector, head-up display with MEMS scanner are all MEMS based computer output interfaces. Additionally, computer mass storage uses a copious amount of MEMS, for example, the hard-disk drive nowadays based on micromachined GMR head and dual stage MEMS micro-actuator. Of course in that last field more innovations are in the labs, and most of them use MEMS as the central reading/writing element.

The optical telecommunication industry has fueled the biggest MEMS R&D effort



so far, when at the turn of the millennium, 10 s of companies started developing optical MEMS switch and similar components. We all know too well that the astounding 2D-switch matrix developed by Optical Micro Machines (OMM) and the 3D-matrix developed in just over 18 months at Lucent are now bed tale stories. However within a few years they placed optical MEMS as a serious contender for the future extension of the optical network, waiting for the next market rebound. Wireless telecommunications are also using more and more MEMS components. MEMS are slowly sipping into handphone replacing discrete elements one by one, RF switch, microphone, filters - until the dream of a 1 mm<sup>3</sup> handphone becomes true (with vocal recognition for numbering of course!). The latest craze seems to be in using accelerometers (again) inside handphone to convert them into game controller, the ubiquitous handphone becoming even more versatile.

Large displays are another consumer product that may prove to become a large market for MEMS. Actually, if plasma and LCD TV seems to become more and more accepted, their price is still very high and recently vendors start offering large display based on MEMS projector at about half the price of their flat panel cousin. Projector based system can be very small and yet provide large size image. Actually, for the crown of the largest size the DLP projecting system from TI is a clear winner as evidenced by the digital cinema theaters that are burgeoning all over the globe. For home theater the jury is still debating - but MEMS will probably get a good share at it and DLP projector and similar technologies won't be limited to PowerPoint presentation.

Finally, it is in space that MEMS are finding an ultimate challenge and some MEMS sensors have already been used in satellites. The development of micro (less than 100kg) and nano (about 10kg) satellites is bringing the mass and volume advantage of MEMS to good use and some project are considering swarms of nano-satellite each replete with micromachined systems.

## 1.4 Mutual benefits between MEMS and microelectronics

The synergies between MEMS development and microelectronics are many. Actually MEMS clearly has its roots in microelectronics, as H. Nathanson at Westinghouse reported in 1967 the “resonant gate transistor” [6], which is now considered to be the first MEMS. This device used the resonant properties of a cantilevered beam acting as the gate of a field-effect transistor to provide electronic filtering with high-Q (see Example 4.5). But even long after this pioneering work, the emphasis on MEMS based on silicon was clearly a result of the vast knowledge on silicon material and on silicon based microfabrication gained by decades of research in microelectronics. Even quite recently the SOI technology developed for ICs has found a new life with MEMS.

But the benefit is not unilateral and the MEMS technology has indirectly paid back this help by nurturing new electronic product. MEMS brought muscle and sight to the electronic brain, enabling a brand new class of embedded system that could sense, think and act while remaining small enough to be placed everywhere. As a more direct benefit, MEMS can also help keep older microelectronics fab running. Actually MEMS most of the times have minimum features size of 1 – 5  $\mu\text{m}$ , allowing the use of older generation IC fabrication equipment that otherwise would just have been dumped. It is even possible to convert a complete plant and Analog Devices has redeveloped an older BiCMOS fabrication unit to successfully produce their renowned smart MEMS accelerometer. Moreover, as we have seen, MEMS component often have small market and although batch fabrication is a must, a large part of the MEMS production is still done using 100 mm (4") and 150 mm (6") wafers - and could use 5-6 years old IC production equipment.

But this does not mean that equipment manufacturer cannot benefit from MEMS. Actually MEMS fabrication has specific needs (deeper etch, double side alignment, wafer bonding, thicker layer...) with a market large enough to support new product line. For example, firms like STS and Alcatel-Adixen producing MEMS deep RIE or EVGroup and Suss for their wafer bonder and double side mask aligner have clearly understood how to adapt their know-how to the MEMS fabrication market.

# Chapter 2

## Introduction to MEMS modeling

### 2.1 Physical scaling laws

The large decrease in size during miniaturization, that in some case can reach 1 or 2 orders of magnitude, has a tremendous impact on the behavior of micro-object when compared to their larger size cousin. We are already aware of some of the most visible implications of miniaturization. Actually nobody will be surprised to see a crumb stick to the rubbed surface of a plastic rod, whereas the whole bread loaf is not. Everybody will tell that it works with the crumb and not with the whole loaf because the crumb is lighter. Actually it is a bit more complicated than that.

The force that is attracting the crumb is the electrostatic force, which is proportional to the amount of charge on the surface of the crumb, which in turn is proportional to its surface. Thus when we shrink the size and go from the loaf to the crumb, we not only decrease the volume and thus the mass but we also decrease the surface and thus the electrostatic force. However, because the surface varies as the square of the dimension and the volume as the cube, this decrease in the force is relatively much smaller than the drop experienced by the mass. Thus finally not only the crumb mass is smaller, but, what is more important, the force acting on it becomes proportionally larger - making the crumb really fly!

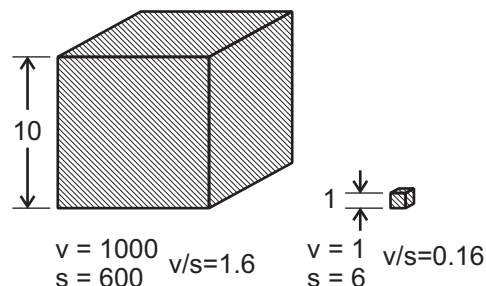


Figure 2.1: Scaling effect on volume, surface and volume/surface ratio.

To get a better understanding, we can refer to Figure 2.1 and consider a cube

whose side goes from a length of 10 to a length of 1. The surface of the bigger cube is  $6 \times 10 \times 10 = 600$  whereas its volume is  $10 \times 10 \times 10 = 1000$ . But now what happen to the scaled down cube? Its surface is  $6 \times 1 \times 1 = 6$  and has been divided by 100 but its volume is  $1 \times 1 \times 1 = 1$  and has been divided by 1000. Thus the volume/surface ratio has also shrunk by a factor of 10, making the surface effect proportionally 10 times larger with the smaller cube than with the bigger one. This decrease of volume/surface ratio has profound implications for the design of MEMS. Actually it means that at a certain level of miniaturization, the surface effect will start to be dominant over the volume effects. For example, friction force (proportional to surface) will become larger than inertia (proportional to mass hence to volume), heat dissipation will become quicker and heat storage reduced: energy storage will become less attractive than energy coupling... This last example is well illustrated by one of the few ever built micromachines, the EMRoS micro-robot from Epson. The EMRoS (Epson Micro Robot System) is not powered with a battery (which stores energy proportional to its volume and becomes less interesting at small scale) but with solar cells whose output is clearly proportional to surface.

Then of course we can dwell into a more elaborate analysis of nature laws and try to see apart from geometrical factor what happens when we shrink the scale? Following an analysis pioneered by W. Trimmer [7], we may describe the way physical quantities vary with scale as a power of an arbitrary scale variable,  $s$ . We have just seen that volume scale as  $s^3$ , surface as  $s^2$  and the volume/surface ratio as  $s^1$ . In the same vein we may have a look at different forces and see how they scale down (Table 2.1).

<b>Force</b>	<b>Scaling law</b>
Surface tension	$s^1$
Electrostatic, Pressure, Muscle	$s^2$
Magnetic	$s^3$
Gravitational	$s^4$

Table 2.1: Scaling of nature forces.

From this table it appears that some forces that are insignificant at large scale becomes predominant at smaller scale. For example lets look at gravity. The Newton's law of gravitation states that the gravity force between two bodies of mass  $m_1$  and  $m_2$  separated by a distance  $r_{12}$  is given by

$$F_g = G_0 \frac{m_1 m_2}{r_{12}^2}$$

where  $G_0$  is the gravity constant. As the scale goes down, the mass depends on volume, thus decrease by  $s^3$ , and the distance obviously by  $s$ , resulting in a scale

dependence of  $s^3 \cdot s^3 / (s^1)^2 \equiv s^4$ . That is, gravity decreases by a factor 10,000 when the scale is shrunk by 10, and thus will be relatively weak at micro-scale. However a more favorable force will be the surface tension force, a line force, which decreases as  $s^1$  making it an important (and often annoying for non-fluidic application) force at micro-scale. The table also reveals that the electrostatic force will become more interesting than the magnetic force as the scale goes down. Of course this simple description is more qualitative than quantitative. Actually if we know that as the size shrinks the electrostatic force will finally exceed the magnetic force, a more detailed analysis is needed to find if it is for the scale of 100  $\mu\text{m}$ , 1  $\mu\text{m}$  or 10 nm. In that particular case it has been shown that the prediction becomes true when the dimensions reach a few  $\mu\text{m}$ , right in the scale of MEMS devices. This has actually been the driver behind the design of the first electrostatic motors by R. Howe and R. Muller[8].

A more surprising consequence of miniaturization is that, contrary to what we would think at first, the relative manufacturing accuracy is sharply decreasing. This was first formalized by M. Madou [10] and it is indeed interesting to see that the relative accuracy of a MEMS device is at a few % not much better than standard masonry. Actually, if it is true that the absolute accuracy of MEMS patterning can reach 1  $\mu\text{m}$  or below, the MEMS size is in the 10  $\mu\text{m}$ -100  $\mu\text{m}$ , meaning a relative patterning accuracy of 1%-10% or even less. We are here very far from single point diamond turning or the manufacturing of large telescope mirror that can both reach a relative accuracy of 0.0001%. So, OK, we have a low relative accuracy, but what does that mean in practice? Let's take two examples for illustrating the issue. First consider a simple door hinge. The pivot of the hinge need to fit very tightly in the barrel to avoid play in the door. With micromachining such hinge will be almost useless because of the large play, compared to the pivot size, it will introduce. Then, let consider the stiffness of a cantilever beam. From solid mechanics the stiffness,  $k$ , depends on the beam cross-section shape and for a rectangular cross-section it is proportional to

$$k = \frac{E h w^3}{4 L^3}, \quad (2.1)$$

where  $E$  is the elasticity modulus,  $h$  is the beam thickness,  $w$  its width and  $L$  its length. For a nominal beam width of 2  $\mu\text{m}$  with an absolute fabrication tolerance of  $\pm 0.2 \mu\text{m}$  the relative accuracy is  $\pm 10\%$ . The stiffness for bending along the width direction varies as a power of 3 of the width and will thus have a relative accuracy of  $\pm 30\%$ . For a stiffness nominal value of 1 N/m, it means that the expected value can be anywhere between 0.7 N/m and 1.3 N/m - the range indicates a potential variation by almost a factor of two! If our design does not support this full range, the yield of the device will be very low. In this particular case, we could improve the relative accuracy figure by taking advantage of the mostly constant absolute fabrication tolerance (here  $\pm 0.2 \mu\text{m}$ ) and increase the beam width. For example, if the beam width grows to 4  $\mu\text{m}$  we reduce the stiffness variation to  $3 \times 4 / 0.2 = 15\%$  - of course it also means doubling the beam length if we want to keep the same

spring constant.

## 2.2 The principles of design and reliability

Since the first days of pressure sensor development, MEMS designers have had to face the complexity of designing MEMS. Actually if IC design relies on an almost complete separation between fabrication process, circuit layout design and packaging, the most successful MEMS have been obtained by developing these three aspects simultaneously (Figure 2.2).

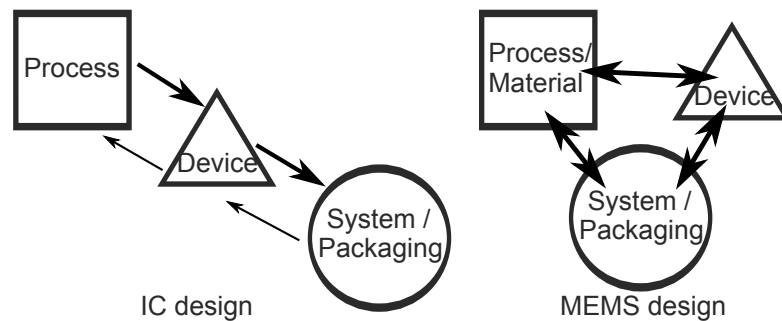


Figure 2.2: IC and MEMS design paradigms.

Actually MEMS fabrication process is so much intertwined with the device operation that MEMS design often involve a good deal of process development. If it is true that some standard processes are proposed by a few foundries (e.g, SOI process, and 3 layer surface micromachining by MemsCap, epitaxy with buried interconnect by Bosch...), there is in MEMS nothing as ubiquitous as the CMOS process.

The success of the device often depends on physics, material property and the choice of fabrication techniques. Actually some industry observers are even claiming that in MEMS the rule is “One Product, One Process” - and many ways to achieve the same goal. Actually we are aware of at least five completely different processes that are currently used to fabricate commercial MEMS accelerometer with about the same characteristics and price - and for at least two companies the accelerometer is their only MEMS product.

And what about packaging then, the traditional back-end process? In MEMS it can account for more than 50% of the final product price and obviously should not be ignored. Actually the designer has to consider the packaging aspect too, and there are horror stories murmured in the industry where products had to be completely redeveloped after trials for packaging went unsuccessful. The main issues solved by MEMS packaging are less related with heat dissipation than with stress, hermetic encapsulation and often chip alignment and positioning. If chip orientation for IC is usually not a concern, it becomes one for single-axis MEMS accelerometer where the chip has to be aligned precisely with respect to the package. This may imply

the use of alignment mark, on the MEMS and in the package. In other case the chip may need to be aligned with external access port. Actually MEMS sensors often need an access hole in the package to bring air or a liquid in contact with the sensing chip, complicating substantially the packaging. One of the innovative approaches to this problem has been to use a first level packaging during the fabrication process to shield the sensitive parts, finally linking the back-end with the front-end. Even for MEMS that do not need access to the environment, packaging can be a complex issue because of stress.

MEMS often use stress sensitive structure to measure the deformation of a member and the additional stress introduced during packaging could affect this function. Motorola solved this problem with its line of pressure sensor by providing calibration of the device after packaging - then any packaging induced drift will be automatically zeroed.

This kind of solution highlights the need to practice design for testing. In the case of Motorola this resulted in adding a few more pins in the package linked to test point to independently tweak variable gain amplifier. This cannot be an afterthought, but need to be taken into consideration early. How will you test your device? At wafer level, chip level or after packaging? MEMS require here again much different answers than ICs.

Understandably it will be difficult to find all the competence needed to solve these problems in one single designer, and good MEMS design will be teamwork with brainstorming sessions, trying to find the best overall solution. MEMS design cannot simply resume to a sequence of optimized answer for each of the individual process, device and packaging tasks - success will only come from a global answer to the complete system problem.

An early misconception about MEMS accelerometer was that these small parts with suspension that were only a few  $\mu\text{m}$  wide would be incredibly fragile and break with the first shock. Of course it wasn't the case, first because silicon is a wonderful mechanical material tougher than steel and then because the shrinking dimension implied a really insignificant mass, and thus very little inertia forces. But sometime people can be stubborn and seldom really understand the predictive nature of the law of physics, preferring to trust their (too) common sense. Analog Devices was facing the hard task to convince the army that their MEMS based accelerometer could be used in military system, but it quickly appeared that it had to be a more direct proof than some equations on a white board. They decided to equip a mortar shell with an accelerometer and a telemetry system, and then fired the shell. During flight, the accelerometer measured a periodic signal, that was later traced back to the natural wobbling of the shell. Then the shell hit his target and exploded. Of course the telemetry system went mum and the sensor was destroyed. However, the 'fragile' sensing part was still found in the debris... and it wasn't broken.

In another example, the DLP chip from Texas Instruments has mirrors supported by torsion hinge  $1\mu\text{m}$  wide and  $60\text{nm}$  thick that clearly seems very prone to failure.

TI engineers knew it wasn't a problem because at this size the slippage between material grains occurring during cyclic deformation is quickly relieved on the hinge surface, and never build-up, avoiding catastrophic failure. But, again, they had to prove their design right in a more direct way. TI submitted the mirrors of many chips through 3 trillions ( $10^{12}$ ) cycles, far more than what is expected from normal operation... and again not a single of the 100 millions tested hinges failed. Of course, some designs will be intrinsically more reliable than others and following a taxonomy introduced by P. McWhorter, at Sandia National Laboratory [11], MEMS can be divided in four classes, with potentially increasing reliability problems.

Class	I	II	III	IV
<b>Type</b>	No moving parts	Moving parts, no rubbing and impacting parts	Moving parts, impacting surfaces	Moving parts, impacting and rubbing surfaces
<b>Example</b>	Accelerometer, Pressure sensor, High-Q inductor, Inkjet nozzle...	Gyroscopes, Resonator, Filter...	TI DLP, Relay, Valve, Pump...	Optical switch, scanner, locking system

Table 2.2: Taxonomy for evaluating MEMS devices reliability

By looking at this table it becomes clearer why developing the Texas Instruments DLP took many more years than developing an accelerometer - the reliability of the final device was an issue and for example, mirrors had originally a tendency to stick to the substrate during operation. TI had to go through a series of major improvements in the material and in the design to increase the reliability of their first design.

## 2.3 MEMS design tools

As we have seen miniaturization science is not always intuitive. What may be true at large scale may become wrong at smaller scale. This translates into an immediate difficulty to design new MEMS structures following gut feeling. Our intuition may be completely wrong and will need to be backed up by models. However simulation of MEMS can become incredibly complex and S. Senturia describes a multi-tiered approach that is more manageable [14] as shown in Figure 2.3.

Some simulation tools like Intellisuite by Intellisense or Coventorware by Coventor have been specifically devised for MEMS. They allow accurate modeling



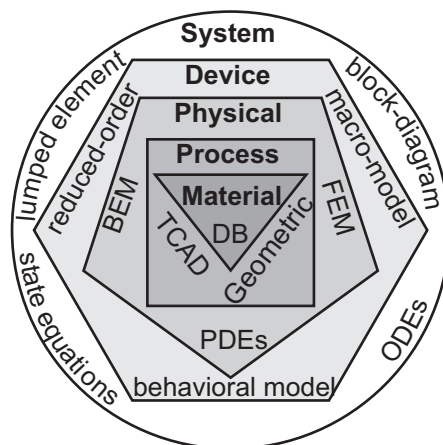


Figure 2.3: MEMS multi-tiered simulation (adapted from [14] and expanded).

using meshing method (FEM, BEM) to solve the partial different equation that describe a device in different physical domains. Moreover, they try to give a complete view of the MEMS design, which, as we said before, is material and process dependent, and thus they give access to material and process libraries. In this way it is possible to build quickly 3D model of MEMS from the mask layout using simulated process. However MEMS process simulation is still in its infancy and the process simulator is used as a simple tool to build quickly the simulation model from purely geometrical consideration, but cannot yet be used to optimize the fabrication process. One exception will be the simulation of anisotropic etching of silicon and some processes modeled for IC development (oxidation, resist development...) where the existing TCAD tools (SUPREM, etc) can be used.

Complete MEMS devices are generally far too complex to be modeled entirely ab initio, and they are first divided into sub-systems that are more manageable. First they can be partitioned between a pure MEMS part and electronics, but they also need to be considered as a mosaic of diverse elements: mechanical structures, actuator, sensor, etc. For example, the MEMS optical switch shown in Figure 2.4 without its control electronics can be divided into a set of optical waveguides, a pair of electrostatic actuator, multiple springs and hinges, a lock and many linkage bars. Then each of the sub-systems can be modeled using reduced order models. For example, behavioral simulation is used with MEMSPro from MemsCap using numerical simulation (ANSYS) to obtain elements characteristics and generate the reduced model, which then is solved in a circuit-analysis software like Spice. Sugar from C. Pister's group at UC Berkeley is also based on behavioral model with discrete elements, but the decomposition of the structure in simpler elements is left to the designer. Still, although the actual tendency is to use numerical modeling extensively, it is our opinion that no good device modeling can be devised without a first analytic model based on algebraic equation. Developing a reduced order model based on some analytic expression help our intuition regains some of its power. For example, seeing that the stiffness varies as the beam width to the

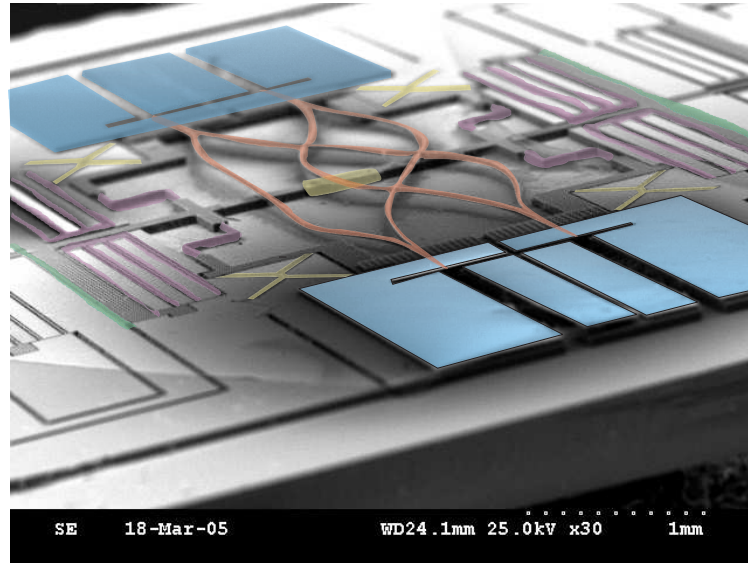


Figure 2.4: MEMS optical switch composed of (red) optical waveguides (yellow) hinges and lock (purple) springs and suspension (green) electrostatic actuator (blue) alignment structure.

cube makes it clearer how we should shrink this beam: if the width is divided by a bit more than two, the stiffness is already ten times smaller. This kind of insight is invaluable. The analytical model devised can also be usefully for validating numerical simulation, which in turn can be used to get detailed insight into the device.

The system level simulation is often not in the hand of the MEMS designer alone, but here block diagram can be used with only a limited set of key state variable. The model may then include the package, the electronics, and the MEMS part will be represented by one or more blocks, reusing the equation derived in the behavioral model.

## 2.4 Lumped elements modeling

The exact internal (hear *physical*) description of a complete microsystem is generally rather complex and time-consuming to fully handle for the designer. Except in the the most simple cases where an analytical solution is existing for the complete structure, modeling the continuous physical system is obtained by discretization. Actually the description of the physical system requires the resolution of the partial differential equations (Maxwell, Newton, ...) of physics by dividing the system in many sub-domains where the equation can be solved analytically or numerically with ease. In this way, the complete continuous problem is replaced by a set of discrete problems, and we intuitively understand that the larger the size of this set is, the closer the resulting discrete model will be from the physical

model. As such, in the FEM method, this approach is systematically applied to divide the continuous system in small sub-blocks, yielding a large number of linear equation that can only be solved numerically with the help of a computer.

However the discretization approach can be used with another strategy: instead of blindly discretizing the complete structure, we can consider only its characteristics of interest and its behaviour. For example, if we have a cantilevered beam submitted to a vertical force at its end as shown in figure 2.5, we can choose to model its behaviour in a number of ways:

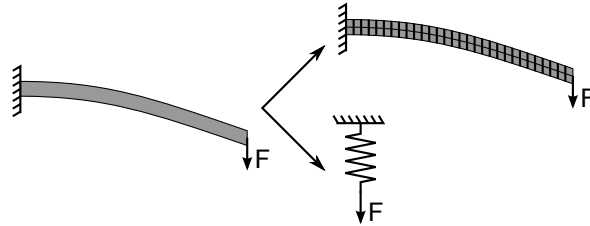


Figure 2.5: Modeling a continuous structure by discrete models: finite elements and lumped element.

- Solve the elastic model with the adequate boundary conditions for the continuous beam (it is here one of these rare cases where the solution is analytic - but it is generally impossible)
- Discretize it with a large number of sub-elements in which the solution of the partial differential equation is trivial (and use a computer to get numerically a complete solution)
- Discretize it with a few lumped elements that capture only the principal characteristics of the beam behavior (a spring extending in the force direction with elastic constant  $k$ )

The last approach loose some of the detail of the device (What happen if the force is sideways? The beam behaves as a spring but it also has a mass, how can we take care of that?) – it is a reduced order model – but it often has the advantage to be amenable to an analytic solution or at least to a simple enough model to be useful during design, even if it requires numerical simulations for refining it.

The block and the circuit representation are two techniques that can be used with lumped elements for complete system simulation, particularly for studying their evolution with time, that is, their dynamic. If these two representations are somewhat equivalent and can be used interchangeably to describe most systems, the circuit elements modeling is usually preferred for sub-systems as it has the advantage to be energetically correct and to intrinsically represent the reciprocal exchange of energy between sub-systems. In the block diagram analysis, we follow state variables representing more abstract variable of interest (the ‘signal’), which

may lose the underlying physics behind the behavior of the elements but becomes simpler to use at system level, particularly to represent change of physical domain (e.g. from electrical to mechanical, or from thermal to mechanical). In Figure 2.6,

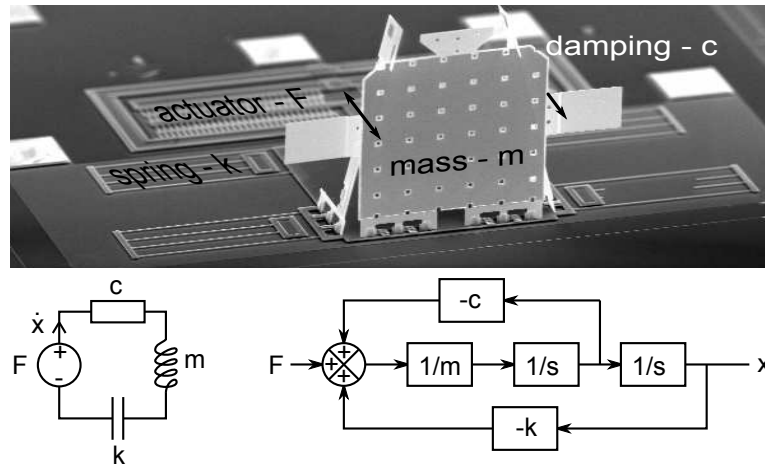


Figure 2.6: Lumped element modeling of a translating micro-mirror using (left) circuit and (right) block analysis.

we illustrate these two techniques by showing how a MEMS device can be analyzed as lumped elements to simulate its behavior. The device is a 3D mirror erected at the edge of a plate which is suspended at four corner over the surface of the substrate and with an electrostatic linear actuator at the back. When the actuator is powered, the mirror can mostly move along one direction ( $x$ ) and we break down the suspended structure in four lumped elements : an actuator with force  $F(t)$ , a spring with elastic constant  $k$ , a mass  $m$ , and a damper mainly linked to the air flow around the structure with damping coefficient  $c$ . The model that we develop is able to answer the simple question : what is the position of the mirror  $x$  as a function of the actuator force  $F(t)$ ? This can be solved by the two models, where the different elements have been connected together, on the left as circuit elements (resistor, capacitor, inductor and voltage source) and on the right as a series of blocks.

We will see soon how these models can be built, but we shall start by using them for the general description of the behavior at system level. The analytical determination of the characteristics of the lumped elements (e.g. what is the elastic constant  $k$  of the spring from geometry and material properties ?) is left to Chapter 4. Actually, it should be noted that for obtaining these characteristics, instead of using analytic formulas, it is generally possible to use numerical physical simulation using *static* model. Then we use the extracted characteristics in the lumped models to quickly obtain the dynamic of the system, a task that would have been very difficult if direct numerical physical *dynamic* simulation would have been performed. This hybrid method, tries to use the best of the analytical and numerical approach and can be very efficient to model complex system, particularly

when the underlying physics is not fully understood by the designer.

### 2.4.1 Block modeling

If artistic drawings, like the beam in Figure 2.5, can be nice for a textbook introduction, it is obvious that the engineer needs another technique to represent the elements inside such system. One method is to use *block diagram*, where each component is represented by a box with an *input* and an *output*. Inside the box we write the *transfer function*, that is the function that relates the input to the output (Figure 2.7). In general the choice of the input and output variable is dic-

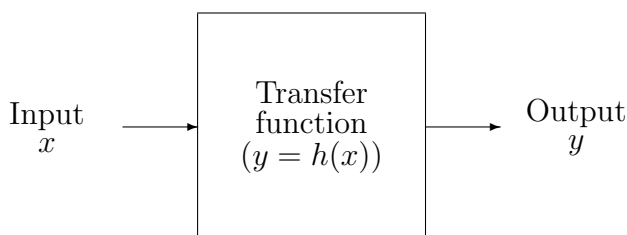


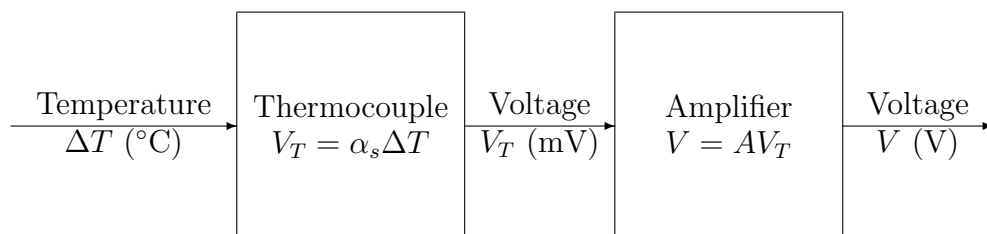
Figure 2.7: Block representation of a system element.

tated by the system, and for a measurement system, as in Example 2.1, the input of the first block should obviously be the measured quantity, the measurand. One of the interest of block analysis is that the input and output variables need not belong to the same energy domain, as it is usually the case in circuit analysis. Thus the block can represent easily complex coupled multiphysics sub-systems, like an electrostatic actuator, simplifying the representation of the complete system.

---

**Example 2.1** Describing a temperature measurement system using block diagram.

WE CONSIDER a thermometer using a thermocouple and an amplifier. The thermocouple deliver a voltage in the mV range that, in a wide temperature range, is linearly proportional to the temperature difference between the two end of the thermocouple (note that as it is a temperature *difference* we may use either Kelvin or degree Celsius). The proportionality constant is  $\alpha_s$  the Seebeck coefficient. This signal is then further amplified by a voltage amplifier of gain  $A$  to give a usable value in the V range.



### 2.4.2 Open and closed-loop systems

In the system theory, the measurement system presented in the Example 2.1, is called an open-loop system. This kind of architecture is typical for measurement system, but *control system* will use closed-loop architecture. The function of such system is to maintain a variable (e.g., temperature, speed, direction...) to a desired value (e.g., 37°C, 1 m/s, 175°...). Sensors are used there to measure the controlled value or a related quantity and actuators to regulate it. Obviously, control systems are everywhere, from the biologic system that adjusts the diameter of the pupil in the eye for controlling the light intensity on the retina, to the lever system allowing automatic leveling of water in the toilet flush!

In control system the output of the system is fed back to the input where an error detector will sense any difference between the desired output and the actual output and act accordingly to correct the error (Figure 2.8). Note that often

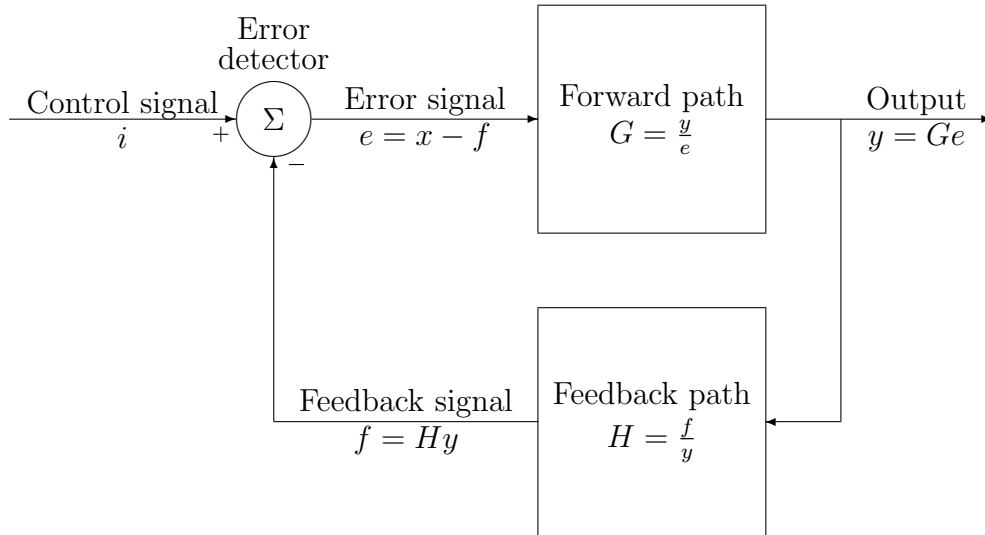


Figure 2.8: A typical control system with a closed-loop architecture (feedback).

‘control system’ are open-loop because they just set the operating value but don’t use any sensor for correcting error. Although they are cheaper to fabricate and simpler to design, they are progressively disappearing in favour of control system with feedback.

The block diagram representation allows to perform simplification and for example the feedback system may be brought to the simpler form of Figure 2.9, where only the input and output signals appear with one single transfer function. Actually for the closed-loop diagram we have the following relationships:

$$\begin{aligned} y &= Ge \\ e &= x - f \\ f &= Hy \end{aligned}$$

We combine the first and second equation

$$y = G(x - f),$$

and then replace the feedback value of  $f$  with its value giving

$$\begin{aligned} y = G(x - Hy) &\Rightarrow y = Gx - GHy \\ &\Rightarrow y(1 + GH) = Gx \\ &\Rightarrow \frac{y}{x} = \frac{G}{1 + GH} \end{aligned}$$

Thus we can redraw the previous block diagram as shown in Figure 2.9. Howe-

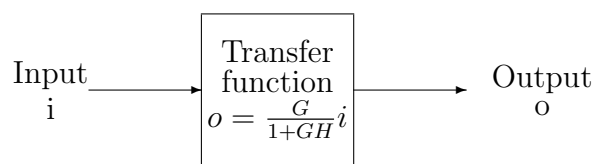


Figure 2.9: Another block diagram representation of a linear closed-loop system.

ver although both diagrams are mathematically equivalent, this new configuration (Fig. 2.9) is slightly less interesting. Actually, now we hide the detail of the implementation too much and it is not clear when we change the sensor what quantity should be changed - whereas it is obvious in the original representation. Thus the use of the simplification should not be abused in block diagram, and each block should try to describe physical element - although computation will need simplification, the use of computer and software like MATLAB<sup>®</sup> with its Simulink<sup>®</sup> module for example, won't require to explicitly perform the simplification.

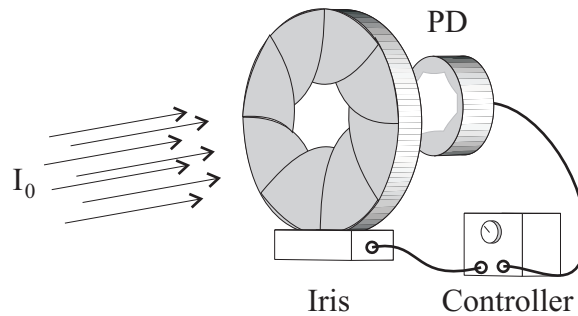
The previous derivation and Example 2.2 have used an implicit hypothesis: the transfer functions  $G$  and  $H$  are simple scalar, allowing to relate input and output of the block by a simple product.

However the transfer functions are usually not simple real constant. For example, in the control system of Example 2.2, the controller used is a simple amplifier, which is called a 'proportional controller'. However, to improve the dynamic characteristics of the control system it would be beneficial to replace the proportional controller by a PID (proportional-integral-derivative) controller, which use the derivative and the integral of the error signal for the actuator. Moreover, we were able to express the effect of the iris by a linear relationship between a voltage and the open area, but what would happen for example for a RLC circuit were the relationship between the voltage and the current is given by a set of integro-differential equations?

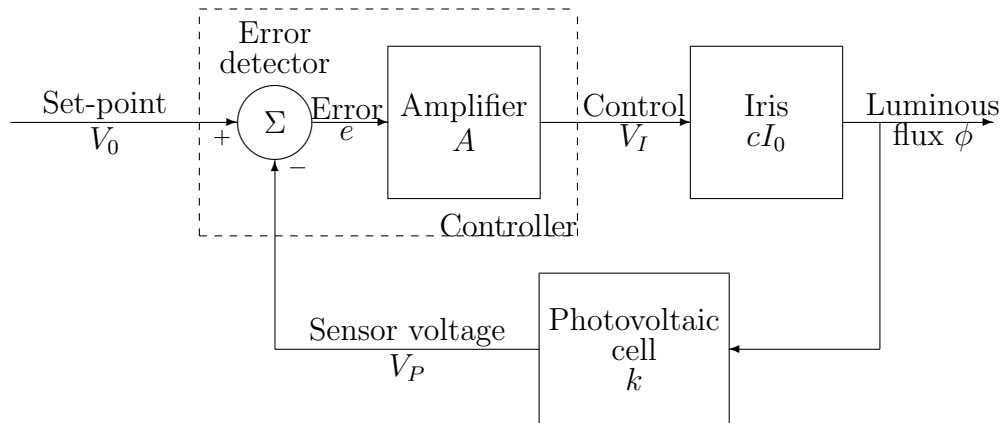
Can we continue to use the simple algebraic manipulation to find the transfer function of the system? In the next section we will see the Laplace's transform that will help answer this question affirmatively!

**Example 2.2** A model of the eye's iris

LET'S CONSIDER a simple model of the eye iris where a photodetector (PD) collects all the light crossing an iris with variable aperture. The photovoltaic cell used as the photodetector delivers a voltage proportional to the radiant flux of the light  $V_P = k\phi$  with  $\phi$  in W.



This voltage is compared in the controller to a reference voltage  $V_0$ , amplified and connected to the iris control, whose aperture  $a$  change linearly with applied voltage ( $a = cV_I$ ). The relationship between the radiant flux and the aperture is given by  $\phi = aI_0$ , where  $I_0$  is the light intensity (in  $\text{W}/\text{m}^2$ ).



The corresponding transfer function for the whole system is given by  $\frac{\phi}{V_0} = \frac{G}{1+GH} = \frac{AcI_0}{1+kAcI_0} \xrightarrow{kAcI_0 \gg 1} 1/k$  which show that the radiant flux on the retina (PD) is independent of the value of the light intensity  $I_0$  if the gain is high,  $kAc \gg 1$ .



### 2.4.3 Linear system and Laplace's transform

Generally the relationship between the output and the input (i.e. the transfer function) is a complex non-linear function of both input and output changing with time. If we restrict ourselves to the case of a single input-single output (SISO) system, this relationship is expressed as,

$$y = f(x, t), \quad (2.2)$$

where  $x$  is the input,  $y$  the output and  $f$  an arbitrary function.

Hopefully, systems (and the blocks describing them) can often be modeled as a *linear* system, which means that the output of the system depends linearly with the input : if the input amplitude is two times larger, the output will also increase by a factor of two. Another consequence is that if we have two signal received simultaneously by the system, the output will be the sum of the output that would have been obtained if the two signal had been measured one at a time. For example, if we measure the light irradiance (or intensity) of two lamps with a linear photodetector, we can either turn on the two lamps and record the reading or turn one lamp after the other and add the two readings together, the result will be the same<sup>1</sup>.

Mathematically these two conditions are represented by:

$$\begin{cases} f(\alpha x, t) = \alpha f(x, t) \\ f(x_1 + x_2, t) = f(x_1, t) + f(x_2, t) \end{cases} \quad (2.3)$$

or equivalently

$$f(\alpha x_1 + \beta x_2, t) = \alpha f(x_1, t) + \beta f(x_2, t) \quad (2.4)$$

Moreover if the system is linear and *time-independent* (i.e., its characteristics don't depend noticeably on time – at least during the time of observation<sup>2</sup>) the relationship between its input and output can be represented by a differential equation,

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \dots + a_1 \frac{dy}{dt} + a_0 y = b_m \frac{d^m x}{dt^m} + b_{m-1} \frac{d^{m-1} x}{dt^{m-1}} + \dots + b_1 \frac{dx}{dt} + b_0 x \quad (2.5)$$

with theoretically<sup>3</sup>  $n > m$ . The sum  $m + n$  is called the order of the system.

The Laplace's transform gives a way to represent the differential equation 2.5 governing a physical system (e.g., the equation of the voltage in an electric circuit,

---

<sup>1</sup>note that the eye is definitely not a linear detector (it follows more or less a logarithmic function) and thus 'visually' two lamps of equal intensity lit together won't create twice as much luminosity (i.e. 'visual' intensity)

<sup>2</sup>it means that we neglect any drift effect, or treat it as a set of quasi-stationary systems

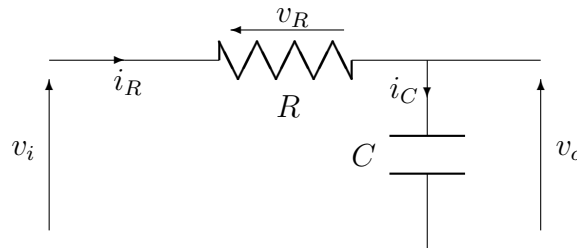
<sup>3</sup>Actually we have used in the previous examples  $n = m$ . It is acceptable in punctual system (i.e., system 'without' spatial extension) as it is the case in lumped model (cf. Appendix B)

---

**Example 2.3** Laplace's transform and block diagram in electric domain.

---

THE RC CIRCUIT shown in the drawing below is also called a low-pass filter. The relationship between the input and output voltage (the two state variables) may be found by looking at the current  $i$  flowing in the circuit.



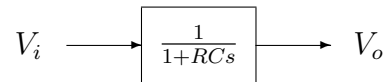
We have  $i_R = v_R/R = (v_i - v_o)/R$  in the resistance, and  $i_C = C dv_o/dt$  for the capacitor. Thus as  $i_R = i_C$ , the relationship between  $v_o$  and  $v_i$  becomes simply:

$$RC \frac{dv_o}{dt} + v_o = v_i$$

Using the table of Laplace's transform properties (Appendix D) we see that the derivation correspond to a product by the Laplace variable  $s$ , and that the sum remains a sum. Thus, in the Laplace's domain the previous equation becomes:

$$RCsV_o + V_o = V_i \Rightarrow \frac{V_o}{V_i} = \frac{1}{1 + RCs} = \frac{1}{1 + \tau s}$$

where  $\tau = RC$  is called the *time constant* of the filter (with dimension of second). Consequently the block diagram *in the Laplace's domain* (note the capital letters) for this element can be represented by:



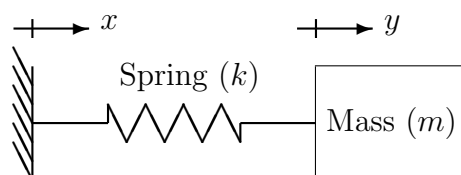
the mechanical response of a suspended mass...) by a simple algebraic equation in  $s$ . Actually, with the Laplace transform, a time derivative becomes a simple product by the variable  $s$  while integration with respect to time becomes roughly a division by  $s$ . Thus, the transfer function for any block in the diagram will keep an algebraic expression, and all the operation and simplification of the ‘block algebra’ will still be valid in the so called “Laplace’s domain”, that is, after the transformation obtained with the Laplace’s transform, provided we have time-independent linear systems.

---

**Example 2.4** Laplace’s transform and block diagram in mechanical domain.

---

ANOTHER example showing the versatility of this method of analysis is found in a simple mechanical system analyzed with lumped elements, a mass and a spring connected together to a frame. Here we will take as state variable the displacement of frame  $x$  (the input) and the displacement of the mass  $y$  (the output).



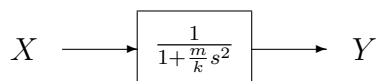
Neglecting the gravity effect, we write down the fundamental equation of dynamic as:

$$F = k(x - y) = m \frac{d^2 y}{dt^2}$$

Using again the table of Laplace’s transform properties (Appendix D) we see that the second order derivative correspond to a product by  $s^2$ . Thus, in the Laplace’s domain we have:

$$k(X - Y) = ms^2 Y \Rightarrow \frac{Y}{X} = \frac{1}{1 + \frac{m}{k} s^2} = \frac{1}{1 + s^2/\omega_n^2}$$

where  $\omega_n = \sqrt{k/m}$  is the natural frequency of the mass-spring system. Consequently the block diagram in the Laplace’s domain for this element is represented by:



Formally speaking the Laplace's transform<sup>4</sup> is given by

$$\mathcal{L}\{f(t)\} = F(s) = \int_0^{\infty} f(t)e^{-st} dt, \quad (2.6)$$

but for the most common functions and operation (like derivation, integration, time delay...) we may use table of transforms (Appendix D) that avoid repeating the integration. We should note that usually the name of a function in the Laplace's domain is written in upper case or block capitals ( $F(s)$ ) and it uses the variable  $s$ .

#### 2.4.4 Analogies and circuit modeling

As we have seen, for block analysis (and Laplace's transform), it doesn't matter what physical effect is behind the system we are considering, because the behavior of electrical, mechanical, thermal or even fluidic systems may be described by similar differential equations. We have already seen in examples 2.3 and 2.4 that electrical and mechanical systems give transfer functions (and differential equation) of similar form. Then by inspecting the equations, we can see that the lumped electrical elements behaves in the same way as the lumped mechanical elements. They are said to be analogue systems.

Accordingly we give simple analogies for system made of lumped mechanical (mass  $m$ , spring  $k$ , damper  $c$ ), electrical (resistor  $R$ , capacitor  $C$ , inductor  $L$ ), thermal (thermal capacitance  $C_\theta$  and resistance  $R_\theta$ ) and hydraulic (hydraulic inductance  $L_h$  and resistance  $R_h$ ) elements in Table 2.3.

Lumped element	Electrical	Mechanical	Thermal	Fluidic
Effort	Voltage	Force	Temperature	Pressure
Flow	Current	Speed	Heat flow	Mass flow rate
Capacitance	$C$	$1/k$	$C_\theta$	$C_h$
Inductance	$L$	$m$		$L_h$
Resistance	$R$	$c$	$R_\theta$	$R_h$

Table 2.3: Example of lumped elements analogies

It is to be noted that this is just a set of example but that the form of the equation may be different if we take other state variables to describe the system

<sup>4</sup>We described here the unilateral Laplace's transform where the integral starts in 0, and thus is defined only for  $t > 0$ . Actually in Engineering all signals are causal - that is have an origin in time - and thus the unilateral Laplace transform takes all its meaning. Bilateral transform will start the integration at  $-\infty$  and allows, for example, for signals that started an infinite time ago, not a real concern for practical case... An advantage of the unilateral transform is that the value of the function at  $t = 0$  can be used (for example when we apply the t-derivative theorem) to set initial condition and solve more complex problems.

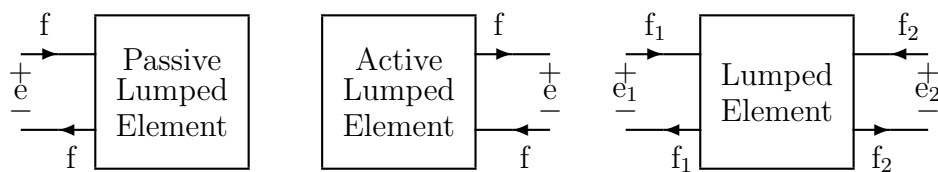
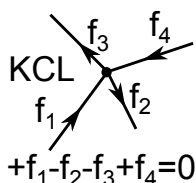
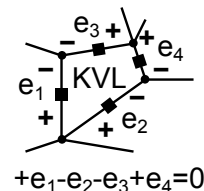


Figure 2.10: One and two port(s) lumped elements in circuit analysis.

(for example we could take the voltage instead of the charge, which will invert the role of the capacitor and inductor) or if the elements are connected differently. The right set of state variable could be dictated by the input and output of the block used but there is a rationale behind the different analogies that makes some more meaningful than other.

Actually, physical variables may be classified as *effort* and *flow* variables. The effort variables follow the Kirchoff's voltage law (KVL), that states that their sum is zero along a closed path, like voltage, pressure, temperature or velocity difference (note that the effort variable has a direction indicated by + and - signs, or sometimes by an arrow placed beneath the circuit path).



The flow variables follow the Kirchoff's current law (KCL), that states that their sum is zero at a any node, like force, current, heat flow, or flow rate (note that the direction of the flow variable is indicated by an arrow on the circuit path). A physical analogy (and not only a mathematical one, only based on the form of the differential equations) will take analog variables inside one class only.

With these analogies in mind, it comes out that instead of using block analysis, another representation of a system can be used: the circuit analysis. In fact, as the internal signal is often an electric signal, it is interesting to represent the rest of the system using a similar description: circuit elements, like capacitor, inductance and resistor. A more fundamental point is that with block analysis we follow the signal in one direction at a time only, whereas circuit analysis intrinsically represents the reciprocal exchange of energy arising in transducers. Actually, if the system receive energy from the environment, its presence will also modify the environment.

Circuit elements will have one port (for effort or flow sources and for passive elements) or two ports (for transformer, gyrator...). A port can be described as a pair of terminals that shares the same flow variable (i.e. current), entering on one terminal and leaving at the other. A different convention is used for passive and active elements, where the flow direction is inverted with respect to the effort across the terminal. A few examples of typical elements are shown in Fig. 2.10.

In this case, the analogies in Table 2.3 are energetically correct. This convention

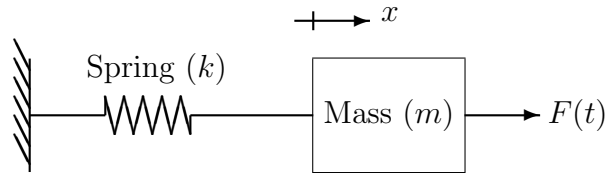
associates the voltage of circuit to all the other effort variables and the current to all the flow variables. Notice that in this case the product of the flow and the effort variable has indeed the dimension of a power (and accordingly, the time integral of the product, the dimension of energy), and thus could allow some ‘sanity check’ during modeling. The main problem encountered with circuit description

---

**Example 2.5** Circuit modeling using analogies.

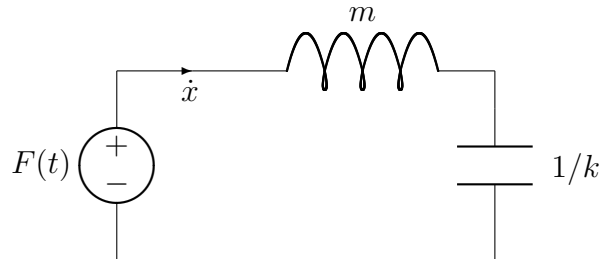
---

**T**AKING again the simple mass-spring mechanical system analyzed previously, but with this time an additional external force  $F(t)$  applied on the mass.



From Table 2.3, we find that the speed ( $\dot{x}$ ) is the flow variable, thus as the mass, the spring end and the external force share the same velocity, they share the same flow and are thus connected in series. Moreover the spring of constant  $k$  becomes a capacitance of value  $1/k$ , the mass  $m$  becomes an inductance of value  $m$  and the external force  $F(t)$  is an effort and it becomes a voltage source.

Thus the equivalent circuit becomes:




---

is: how to connect the lumped elements together? This is solved by understanding that elements that share the same effort are connected in parallel, while elements sharing the same flow are connected in series.

We note that when we have the circuit-element model of our sub-system, instead of deriving the differential equation we may use impedances and Kirchoff's laws to directly establish transfer function. Actually, in the Laplace domain, for each passive circuit element the relationship between the voltage drop across one port  $U$  (the effort) and the current  $I$  (the flow) can be written as :

$$U = Z_i I$$

where  $Z_i$  is the impedance of the element  $i$ .

For resistance we have,

$$Z_R = R$$

for capacitance,

$$Z_C = \frac{1}{Cs}$$

and for inductance,

$$Z_L = Ls.$$

Then by applying Kirchoff's Voltage and Current laws as seen in Example 2.6 it becomes very simple to find the relationship between any voltage (effort) or current (flow) variable in the sub-system and particularly, to establish its transfer function.

---

**Example 2.6** Establishing filter transfer function using complex impedance.

---

WE REVISIT the case of Example 2.3 by using complex impedance. Actually we have in Laplace domain:

$$V_o = V_C = \frac{I_C}{Cs}$$

moreover, in the circuit, applying Kirchoff's voltage law we can write:

$$-v_i + v_R + v_o = 0$$

thus in Laplace's domain,

$$V_i = RI_R + V_o \Rightarrow I_R = \frac{V_i - V_o}{R}$$

which can be replaced in the first equation as  $I_R = I_C$ ,

$$V_o = \frac{V_i - V_o}{R} \frac{1}{Cs}$$

giving the expected final relationship in Laplace domain without any differential equation:

$$V_o = \frac{V_i}{1 + RCs}$$


---

## 2.5 Dynamic analysis

In principle to obtain the dynamic response of a system (that is, its evolution with time) we need to solve the differential equation that describes the system (eq. 2.5) with the appropriate initial conditions.

However with the lumped element analysis using either block or circuit elements we have simpler ways to deal with this problem<sup>5</sup>. Actually, we will be using again

---

<sup>5</sup>It could be noted that the lumped elements we use in our modeling are actually punctual

the Laplace's transform for studying the system dynamic in the time domain, or a very close formalism, the Fourier transform, to study it in the frequency domain. Actually, the dynamic study of a system can be accurately performed both in the time domain (where the variable is  $t$ ) or in the frequency domain (where the variable is  $f$  or  $\omega$ ).

### 2.5.1 Time domain analysis with Laplace's transform

With the Laplace's transform the tedious task of solving the differential equation<sup>6</sup> is replaced by a lookup in a table of transforms (cf. Appendix D.2) and algebraic manipulations facilitated by simple properties (cf. Appendix D.1).

The procedure can be summarized as follow:

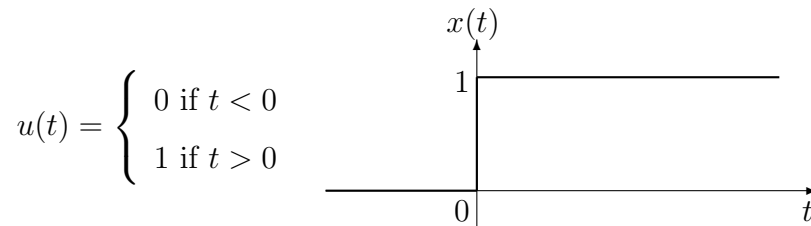
**First step** using the properties of the Laplace's transform (Table D.1), transpose the integro-differential equation of the system ( $t$  variable) to an equation in the Laplace's domain ( $s$  variable), to obtain the transfer function of the system;

**Second step** Transpose the input signal in the Laplace's domain using the table of functional transform (Table D.2);

**Third step** Compute the output signal in the Laplace's domain by multiplying the transfer function with the input signal;

**Fourth step** Transform back this subsidiary solution to the time domain using a table of inverse Laplace's transforms (Table D.2). Before looking-up in the table of inverse transform, it is better to perform some algebraic manipulation on the equation in  $s$  to bring it to a form consisting of a sum of 'simple' polynomial, whose inverse transforms are easily found in the table.

Although this procedure works for any input signal, we usually limit ourselves to a few classes of typical signals. The most useful one is the step signal, that represents a sudden change in the input, and which is represented by a step function:




---

elements: we neglect their spatial extension and accordingly the time for the signal to travel through one element is completely ignored. The dependence of the response of a system with time is of another nature, linked with the existence of energy storage elements that requires time to be charged or discharged.

<sup>6</sup>Laplace's transform method works both for ordinary differential equations (ODE) and partial differential equation (PDE), but we will see here only the simpler problems, described by the former.



As we can see from the table D.2, the Laplace's transform of this function is very simple,

$$U(s) = \mathcal{L}(u(t)) = \frac{1}{s}.$$

The response of a system to the step signal is simply called the *step response* as in Example 2.7, and it gives a good insight on how quickly the system will react when there is a sudden change in the input (e.g., for an accelerometer it would mean when a car crashes and the acceleration increases quickly).

## 2.5.2 Frequency domain analysis with Fourier's transform

However, the time domain analysis is generally not sufficient to get a good understanding of the properties of a system, and often the dynamic analysis is completed by an analysis in the frequency domain. In frequency domain, the plot of a function versus frequency is called the *spectrum* of the function and we talk about the spectrum of a signal or the spectrum of a transfer function.

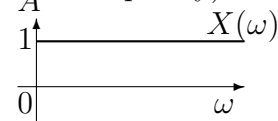
To obtain the spectrum of the response, we need to compute the Fourier transforms of the transfer function and of the input signal. These computations are again obtained easily with the Laplace's transform. The trick is even simpler than for the time domain analysis<sup>7</sup> : we need to replace in the function in the Laplace's domain the  $s$  variable with  $j\omega$ , where  $j$  is the complex imaginary unit (i.e.,  $j^2 = -1$ ) and  $\omega$  is the angular frequency of the signal (i.e.,  $\omega = 2\pi f$  where  $f$  is the frequency).

For any value of the frequency, the transfer function becomes a complex number ( $H(\omega) = A(\omega)e^{j\phi(\omega)}$ ) whose amplitude  $A(\omega)$  is the gain (or amplitude ratio) and whose phase  $\phi(\omega)$  is the phase-shift induced by the transfer function.

Then, as we did in the Laplace's domain, we can obtain the spectrum of the response to any input signal by multiplying the transfer function in frequency domain by the spectrum of the input signal. The spectrum of the input signal is again obtained by replacing the  $s$  variable with  $j\omega$  in the Laplace's transform of the signal<sup>8</sup>. The resulting spectrum of the input signal is a complex quantity with an amplitude and a phase and it can be represented by its amplitude spectrum (the amplitude vs frequency) and its phase spectrum (the phase vs frequency).

Again, in the frequency domain there is a particular response that is more often studied, which is called the *sinusoidal steady state* response. The sinusoidal steady state response is the response of the system when the input signal is an infinite sum of sinusoid of all frequencies

with equal amplitude and phase and after we have waited long enough for the transient response to disappear. This particular input signal has thus a constant amplitude spectrum over all the frequencies and a constant phase spectrum of 0



<sup>7</sup>Actually what we compute here is the Fourier's transform, that could be obtained directly with integral or tables too - but as we already have the Laplace's transform this is a much simpler way. Note that the transform is a complex quantity and a refresher in complex numbers is provided in Appendix E

<sup>8</sup>Actually we are again performing a Fourier's transform of the input signal.

**Example 2.7** Response to a step signal for a low-pass filter

LET'S TRY to derive the output voltage of the previous low-pass filter when the input voltage brutally change from 0 V to 10 V. We have already computed in the previous example the Laplace's transform of the filter transfer function, and we have found:

$$H(s) = \frac{V_o}{V_i} = \frac{1}{1 + \tau s}$$

We also need to transform in the Laplace's space the input signal. This signal is a step function with amplitude 10 V, that is thus represented by,  $v_i(t) = 10u(t)$ , where  $u(t)$  is the usual name for the step function (i.e.,  $u(t) = 0$  for  $t < 0$  and  $u(t) = 1$  for  $t \geq 0$ ). We use the properties of table D.1 and then look in the table D.2 to find that :

$$V_i(s) = \mathcal{L}\{v_i(t)\} = \mathcal{L}\{10u(t)\} = 10\mathcal{L}\{u(t)\} = \frac{10}{s}$$

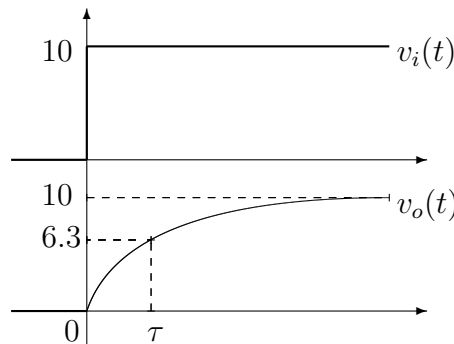
We compute the output voltage using simple algebra :

$$V_o(s) = H(s)V_i(s) = \frac{1}{1 + \tau s} \frac{10}{s} = \frac{10}{s(1 + \tau s)}$$

Finally, we just need to transform back the solution to time domain:

$$\begin{aligned} V_o(s) &= \frac{10}{s(1 + \tau s)} = 10 \frac{1/\tau}{s(1/\tau + s)} = 10 \left( \frac{1}{s} - \frac{1}{1/\tau + s} \right) \\ \Rightarrow v_o(t) &= \mathcal{L}^{-1}\{V_o(s)\} = \mathcal{L}^{-1} \left\{ 10 \left( \frac{1}{s} - \frac{1}{1/\tau + s} \right) \right\} \\ &= 10 \left( \mathcal{L}^{-1} \left\{ \frac{1}{s} \right\} - \mathcal{L}^{-1} \left\{ \frac{1}{1/\tau + s} \right\} \right) \\ \Rightarrow v_o(t) &\stackrel{t \geq 0}{=} 10(1 - e^{-t/\tau}) = 10u(t)(1 - e^{-t/\tau}) \end{aligned}$$

The output of the circuit will never reach the final value but rises slowly towards it asymptotically (we have 63% of the max value when  $t = \tau$ ).



(see the amplitude spectrum in the inset<sup>9</sup>). To obtain the corresponding response we need, as we just said, to multiply the transfer function in frequency domain by the input signal spectrum – it is of course trivial because the spectrum of the input signal is constant and equal to 1. Thus the sinusoidal steady state is actually simply  $H(j\omega)$ , that is, the Fourier's transform of the transfer function. We note here that as the analysis is actually performed in frequency domain we don't need to 'transform back' the result as we did for the time domain analysis with the Laplace's transform.

---

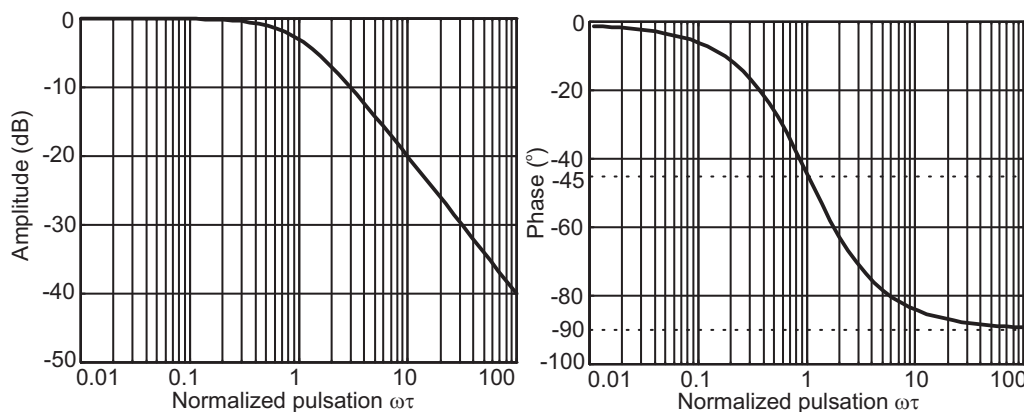
**Example 2.8** Sinusoidal steady state response of a low-pass filter

---

LET'S LOOK again at the low-pass filter. We have already computed the Laplace's transform of the filter transfer function, that is  $H(s) = \frac{V_o}{V_i} = \frac{1}{1+\tau s}$ . To get the transfer function in the frequency domain we replace  $s$  by  $j\omega$  and get :

$$H(j\omega) = \frac{1}{1+j\tau\omega} = \frac{1}{1+\tau^2\omega^2} - j\frac{\tau\omega}{1+\tau^2\omega^2} = \frac{1}{\sqrt{1+\tau^2\omega^2}} e^{-j\arctan\tau\omega}$$

We have here extracted the amplitude (modulus) of the complex transfer function and its phase using the properties of complex numbers (cf. Appendix E).



Looking at the amplitude transfer function, it becomes obvious why such a circuit is called a low-pass filter: it let the signal with low-frequency pass ( $\omega\tau < 1$  or  $f < 1/(2\pi RC)$ ) and attenuates substantially the signal with higher frequency.

---

The sinusoidal steady state response ( $H(\omega) = A(\omega)e^{j\phi(\omega)}$ ) is a complex quantity, with an amplitude and a phase and it is customary to plot it as a function of frequency in a pair of plots called the Bode diagram. On one plot, the amplitude is plotted in decibel ( $20 \log A(\omega)$ ) and on the other the phase ( $\phi(\omega)$ ) is plotted in degree. In both plots the horizontal axis is the frequency on a logarithmic scale,

---

<sup>9</sup>How does such a signal with a constant frequency spectrum and phase look like? In the time domain it is a pulse, that is a signal of very short duration and finite energy  $\perp$ . Mathematically, it is represented by the dirac function  $\delta(t)$  whose Fourier's transform is of course... 1 – constant over the whole frequency spectrum – as confirmed by Laplace's transform table.

allowing display of a wide range of frequency (e.g., from 0.1 Hz to 1 MHz)<sup>10</sup>.

At the frequencies where the amplitude (or gain) spectrum is larger than 1, the input signal is said to be amplified by the system, while when it is smaller than 1 it is attenuated. The phase of the sinusoidal steady state response is also a very important parameter of a system. For example, it is used to determine if a closed-loop system is stable or not, that is, if it will start oscillating by itself or saturate<sup>11</sup>.

Even with the Laplace's transform method, a general linear system of order  $m+n$  is quite tedious to handle. It is thus advisable to try to simplify the problem by lowering the order of the system using valid approximation.

If we remember Example 2.2, we have already used a crude approximation for the photodetector transfer function by using a constant ( $V_P = k\phi$ ), corresponding actually to a model of 'zero' order. It is valid only for slowly varying signals, where the output can be considered to match the input 'instantaneously' (i.e., at a much larger speed than the time it take for the signal to change). Actually strictly speaking, a 'zero' order transfer function is not physical and a less crude model for the photodiode will be using a first order model.

Actually the transfer function of many microsystems can be quite accurately described by an equation of first (i.e,  $m+n = 1$ ) or second ( $m+n = 2$ ) order. This happen even in complex system because there is often only one or two elements dominating its behavior. Of course, this is just an approximation and it will be generally valid only for a limited frequency range, but it is important for obtaining an insight in the behavior of complex systems and justify to study a bit more these two types of systems.

### 2.5.3 First-order model

Typical example of microsystems that may be modeled using a first-order approximation are thermal sensors, certain chemical sensors, a photodiode, a seismic sensor... and we have already seen a first-order transfer function in the example 2.7. In all this cases the inertia of the system (i.e., its ability to resist change) is much larger than any other characteristics. Then, the relationship between the

---

<sup>10</sup>Another often used representation of the sinusoidal steady state response is the Nyquist diagram, where instead of plotting the amplitude and phase on two separate diagrams, we use a parametric plot ( $f$  is the parameter) of the complex frequency response with the real part on the X-axis and the imaginary part on the Y-axis.

<sup>11</sup>Control theory tells that a closed-loop system is not stable if the signal fed back to the input has been amplified (i.e.,  $A \geq 1$ ) and is in opposition of phase (i.e.,  $\phi = 180^\circ$ ) - such condition can be observed directly on the Bode diagram. The very important problem of stability in closed-loop system is beyond the scope of this course, but interested reader may refer to control theory books.

input ( $x$ ) and the output ( $y$ ) is given in the time and the Laplace's domain by :

$$y = Gx - \tau \frac{dy}{dt} \quad (2.7)$$

$$Y(s) = \frac{G}{1 + \tau s} X(s) \quad (2.8)$$

where  $G$  is the static gain of the system (e.g., the sensitivity of a sensor or 1 in the low-pass filter example) and  $\tau$  is its time constant. The static gain is given by the ratio between the input and the output after an 'infinite' time has elapsed after the input has changed - that is when the transient part of the signal at the output has died out and  $dy/dt = 0$ . It is also simply the value of the transfer function when the frequency of the input is 0 ( $G = H(0)$ ).

### 2.5.3.1 Step-response

The step response of the first order system is given by:

$$y_{\text{step}}(t) = Gu(t)(1 - e^{-t/\tau}) \quad (2.9)$$

By observing the step response of the first order system in Figure 2.11 it appears that waiting about  $> 7\tau$  is sufficient to obtain about 99.9% of the final response (i.e.,  $1 - e^{-7}$ ).

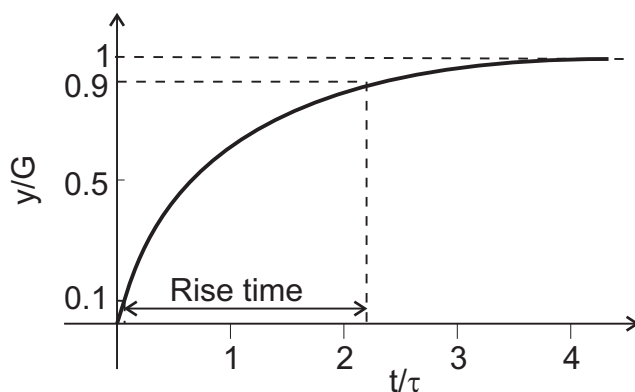


Figure 2.11: Step response ( $x(t) = u(t)$ ) of a first order system.

This behavior will fix the speed of the system and thus will ultimately limit the capability to follow the input at a high rate.

From the measured step-response of a system, we can retrieve the value of the two parameters of the first-order model, the static gain and the time constant. The static gain ( $G$ ) is obtained as the value of the gain (the ration between the output and the input) when the 'transient' response has vanished, that is, after a 'long enough' time. The time constant  $\tau$  can be obtained as the time for the output to reach 63.5% of its maximum value (i.e.,  $1 - e^{-1}$ ). A more practical definition

for this quantity (avoiding difficulty to define signal level near the start and end points exactly) uses the rise time of the system, which is defined as the time to go from the 10% level to the 90% level. For first order system we have  $\tau = t_{\text{rise}}/2.35$ .

### 2.5.3.2 Frequency response

The transfer function in the frequency domain of the first order system is given by:

$$H(j\omega) = \frac{G}{\sqrt{1 + \tau^2\omega^2}} e^{-j \arctan \tau\omega} = A(\omega) e^{j\phi(\omega)} \quad (2.10)$$

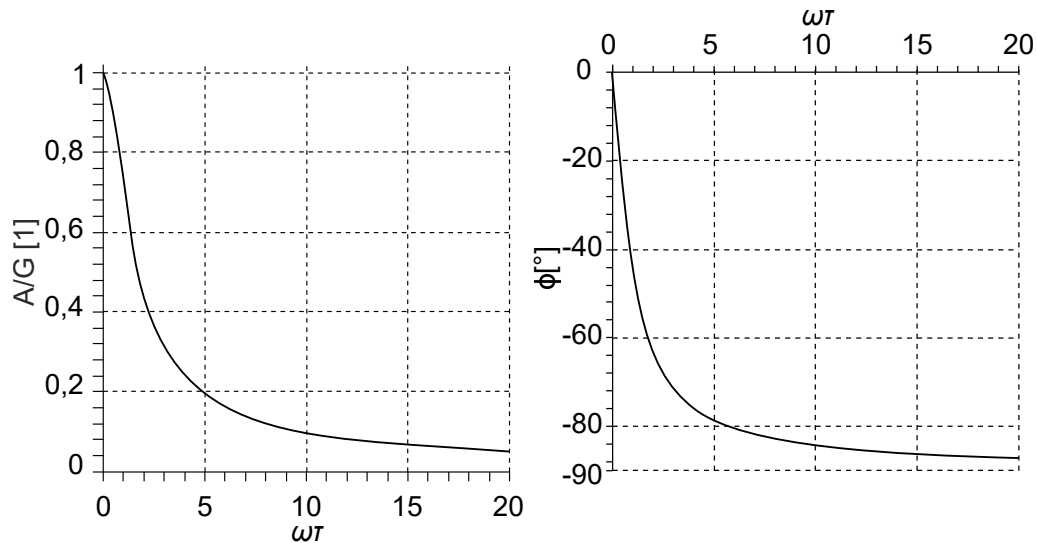


Figure 2.12: Plot of a first order transfer function (linear frequency scale)

Actually we have already seen this response in Example 2.8, but instead of reproducing the Bode's plot here (with its logarithmic scales for frequency and amplitude), we show in Figure 2.12 a linear plot, without using dB nor logarithmic frequency scale. Still it should be understood that the two figures show exactly the same data – there is only a scale difference! The transfer function amplitude remains almost constant when  $\omega < 1/\tau$ . When  $\omega = 1/\tau$ , the gain has been divided by  $\sqrt{2}$  (or decrease by about  $20 \log(\sqrt{2}) \approx -3\text{dB}$  on the log plot). This frequency  $f_c = 1/2\pi\tau$ , is called the *cut-off frequency* and at  $f_c$  the phase shift is exactly  $45^\circ (\pi/4)$ , which provide another mean to obtain the time constant of the circuit.

For a frequency much larger than the cut-off frequency (i.e.,  $f \gg f_c$  or  $\omega\tau \gg 1$ ), the transfer function amplitude becomes simply  $A(\omega) \approx G/(\omega\tau)$ . Thus, when the frequency is multiplied by a certain factor, the signal is divided by the same factor. For example the signal is divided by 10 when the frequency is multiplied by 10 (i.e., on the log plot, the amplitude decreases by 20 dB per decade) or it is divided by 2 when the frequency is multiplied by 2 (that is  $\approx -6$  dB per octave on the log

plot). In this frequency range the transmittivity (corresponding to the amplitude of the transfer function) changes too rapidly usually preventing to use the system predictably.

### 2.5.4 Second-order model

The next simplest model after the first-order model is the second-order model where we have  $m + n = 2$ . Such model describes systems where there are two inertial components of similar magnitude coupled together, or a transfer between kinetic energy and potential energy. Typical examples are (R)LC circuits, some thermal sensors, most of the mechanical sensors (accelerometer, pressure sensor, etc) or linear actuators... The general equation governing such system is given in time and Laplace's domain by :

$$y = Gx - 2\frac{\zeta}{\omega_n} \frac{dy}{dt} - \frac{1}{\omega_n^2} \frac{d^2y}{dt^2} \quad (2.11)$$

$$Y(s) = \frac{G\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} X(s) \quad (2.12)$$

where  $G$  is the static gain,  $\omega_n$  the natural frequency and  $\zeta$  (pronounced 'zeta') the damping ratio.

Using analogies, Table 2.4 shows the parameters of first and second order systems expressed using typical lumped elements existing in the different physical domains.

Characteristic	Mechanical	Electrical	Thermal
$\omega_n$	$\sqrt{\frac{k}{m}}$	$\sqrt{\frac{1}{LC}}$	
$\zeta$	$\frac{c}{2\sqrt{km}}$	$\frac{R\sqrt{C}}{2\sqrt{L}}$	
$G$	$1/k$	$C$	
$\tau$	$\frac{c}{k}4$	$RC$	$R_t C_t$

Table 2.4: First and second order analog model.

#### 2.5.4.1 Step-response

The step response of a second-order system is more complicated than for a first-order one, and depends on the value of  $\zeta$ .

$$y_{\text{step}}(t) = Gu(t) \left\{ 1 - e^{-\zeta\omega_n t} \left[ \cosh \left( \sqrt{\zeta^2 - 1} \omega_n t \right) + \frac{\zeta}{\sqrt{\zeta^2 - 1}} \sinh \left( \sqrt{\zeta^2 - 1} \omega_n t \right) \right] \right\} \quad (2.13)$$

The complicated expression does not give much insight, but we can study limiting cases depending on the value of  $\zeta$ . When the damping ratio is very small (low damping) we have,

$$\zeta \ll 1 \Rightarrow \sqrt{\zeta^2 - 1} \approx j \Rightarrow y_{\text{step}}(t) = Gu(t) \{1 - e^{-\zeta\omega_n t} \cos(\omega_n t)\}$$

that is an oscillation of decreasing amplitude. In another extreme case, we consider a large damping and we have

$$\zeta \gg 1 \Rightarrow \sqrt{\zeta^2 - 1} \approx \zeta \Rightarrow y_{\text{step}}(t) = Gu(t) \{1 - e^{-\zeta\omega_n t}\}$$

representing a slowly increasing response, similar to first order response. In the middle we have the case  $\zeta = 1$ , where we get<sup>12</sup>:

$$\zeta = 1 \Rightarrow \sqrt{\zeta^2 - 1} = 0 \Rightarrow y_{\text{step}}(t) = Gu(t) \{1 - e^{-\omega_n t}(1 + \omega_n t)\}$$

This last case gives the fastest possible settling time without any overshoot or oscillation. These three cases are schematically depicted in Figure 2.13. We show

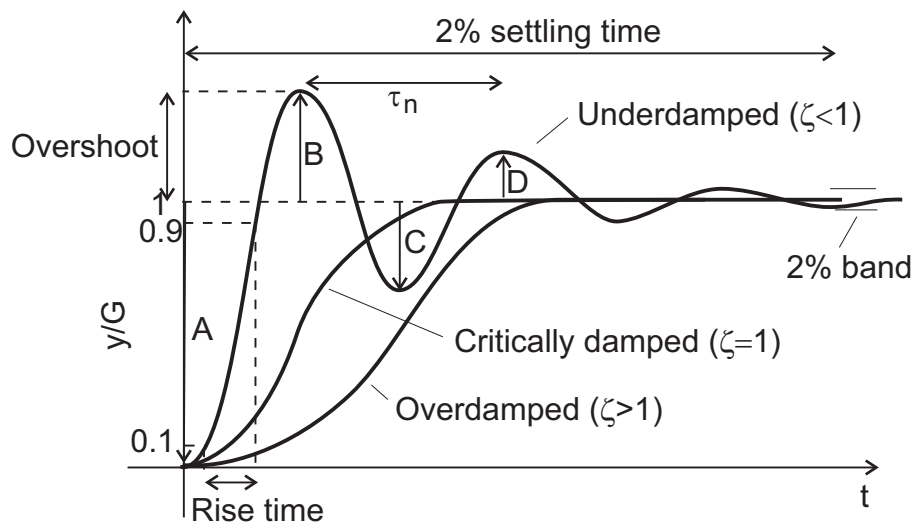


Figure 2.13: Step response for a 2nd order transfer function with different damping ratio  $\zeta < 1$  (under-damped),  $\zeta = 1$  (critically damped), and  $\zeta > 1$  (over-damped).

in the Figure the typical terms used to describe the response: overshoot, settling time and rise time. We have figured also the case where the oscillations and the overshoot just disappear which is called a critically damped system. This happen when the term  $\sqrt{\zeta^2 - 1}$  cease to be complex and becomes real, that is when  $\zeta \geq 1$ . Note that if it is possible to tolerate overshoot, the rise time may be shortened by using a slightly under-damped system ( $\zeta < 1$ ). Usually over-damped systems

<sup>12</sup>The simplest at this stage to obtain the function is to see that  $\lim_{x \rightarrow 0} \sinh(ax)/x = a$  using limited development of the function close to 0



( $\zeta > 1$ ) are avoided, because they don't bring advantage over the two other cases (still, a slightly over-damped system  $\zeta \gtrsim 1$  may be useful to improve the robustness of the system).

It can be noted that in an under-damped ( $\zeta < 1$ ) second order system, we can easily retrieve different parameters of the model with the step response. The static gain ( $G$ ) is again obtained as the value of the gain (output/input) when the 'transient' response has vanished, that is after a long enough time. Then, the natural frequency ( $\omega_n$ ) is roughly obtained from the period  $\tau_n$  of the oscillations ( $\cos \omega_n t$  term) as

$$\omega_n = \frac{2\pi}{\tau_n}.$$

Finally, the damping ratio is obtained by taking the ln of the ratio of the amplitude of two consecutive oscillations. Actually the amplitude of the oscillation decreases as  $e^{-\zeta \omega_n t}$ , thus we have  $\ln(A/B) = \ln(B/C) = \ln(A/C)/2 = \pi\zeta/\sqrt{1-\zeta^2} \stackrel{\zeta < 0.3}{\approx} \pi\zeta$ , yielding :

$$\zeta \approx \ln(A/B)/\pi \text{ or } \zeta \approx \ln(B/C)/\pi \text{ or } \zeta \approx \ln(A/C)/2\pi \text{ or } \dots$$

This first approximate expression is particularly useful because it needs only the relative overshoot ( $B/A$ ) (e.g., if the overshoot is 40% of the final amplitude we have  $\zeta \approx -\ln(0.4)/\pi \approx 0.29$ ). Note that if the overshoot is smaller than about 40% we should not use the approximate equation ( $\zeta \approx \ln(A/B)/\pi$ ) but use the exact expression with the square root given above.

### 2.5.4.2 Frequency response

For the frequency response, the equation is again much more involved than for a first order model, but the Bode plot will reveal its most important features. Again we replace  $s$  by  $j\omega$  in the Laplace's expression of the transfer function, and separate the amplitude and phase part of the complex expression.

$$\begin{aligned} H(j\omega) &= \frac{G\omega_n^2}{-\omega^2 + 2j\zeta\omega_n\omega + \omega_n^2} \\ &= \frac{G}{\sqrt{(1 - \omega^2/\omega_n^2)^2 + (2\zeta\omega/\omega_n)^2}} e^{-j \arctan \frac{2\zeta\omega/\omega_n}{1 - (\omega/\omega_n)^2}} \end{aligned} \quad (2.14)$$

Using these formulas we may plot the Bode diagram using the MATLAB code given in Annex G. In the plot of Figure 2.14 we have used the normalized pulsation  $\omega/\omega_n$ , a gain of 1 and we have varied the damping ratio between 0.05 and 5. The main feature of the amplitude plot is the resonance phenomena that appears for certain values of the damping ratio. We observe a marked increase of the transfer function amplitude that appears around the natural frequency, and where the output becomes larger than the input.

The frequency where the transfer function has its maximum is called the resonance frequency and can be found by looking for the maximum of the amplitude

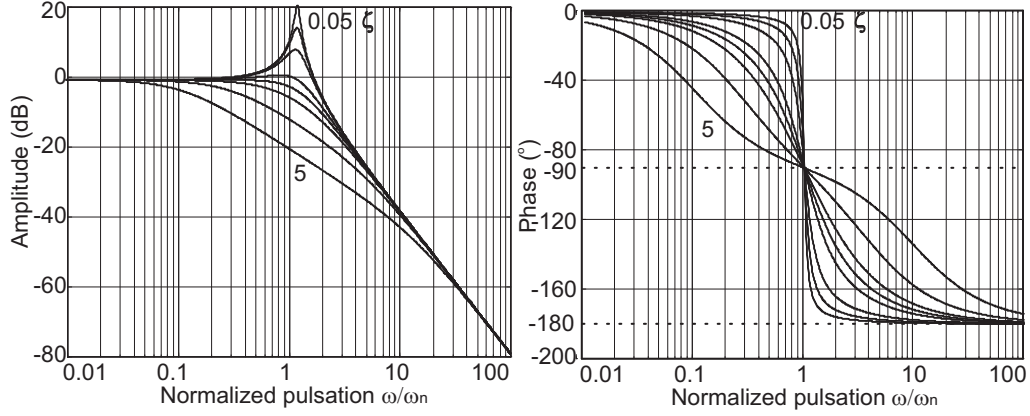


Figure 2.14: Bode diagram for a 2nd order transfer function with different damping ratio  $\zeta = [0.05, 0.1, 0.3, 0.5, 1, 3, 5]$ .

term in eq. 2.14. This maximum is attained when the denominator of the amplitude reaches a minimum, thus by equating the derivative of  $\sqrt{(1 - \omega^2/\omega_n^2)^2 + (2\zeta\omega/\omega_n)^2}$  with 0 and we find that the resonance frequency is given by :

$$\omega_0 = \omega_n \sqrt{1 - 2\zeta^2} \text{ for } \zeta < 1/\sqrt{2}. \quad (2.15)$$

At the same time we find that this maximum amplitude reached by the output when  $\omega = \omega_0$  is  $y_{\max} = Gx_{\max}/(2\zeta\sqrt{1 - \zeta^2})$  (or simply  $y_{\max} = Gx_{\max}/(2\zeta)$  when  $\zeta \lesssim 0.3$ ) : the amplitude at resonance is multiplied by a factor  $1/2\zeta$ . This may be used to amplify the response of a system - but it is rarely used as it only works for a limited band of frequency and is hard to control.

The condition that  $\zeta < 1/\sqrt{2}$  in eq. 2.15 implies that for values larger than  $1/\sqrt{2}$  there is no resonance : the amplitude does not increase above the value at  $\omega = 0$ . This can be easily proved, by using the value of  $\omega_0$  in the amplitude factor of eq. 2.14, and by searching for which value of  $\zeta$  the resonance (i.e., the amplification) disappears. We have:

$$\begin{aligned} \frac{G}{\sqrt{(1 - \omega_0^2/\omega_n^2)^2 + (2\zeta\omega_0/\omega_n)^2}} &> G \\ \Rightarrow \left(1 - \sqrt{1 - 2\zeta^2}\right)^2 + \left(2\zeta\sqrt{1 - 2\zeta^2}\right)^2 &< 1 \\ \Rightarrow -4\zeta^4 + 4\zeta^2 &< 1 \\ \Rightarrow \zeta &< 1/\sqrt{2} \end{aligned}$$

We can note that for  $1/\sqrt{2} < \zeta < 1$  there is no resonance in the frequency response whereas the step response clearly shows an overshoot - although these two phenomena are interrelated, there is no direct relationship. This shows that we need to take care of too quick conclusions when we try to deduce the frequency response from the step response (or the reverse), and that their careful individual study is interesting!

---

**Example 2.9** Experimental determination of the parameter of a second order model.

---

INTERESTINGLY, if we measure and draw the Bode plot of a system that could be modeled as a second order system, we may be able to easily find all the characteristics ( $\omega_0$ ,  $\zeta$ , and  $G$ ) of the model.

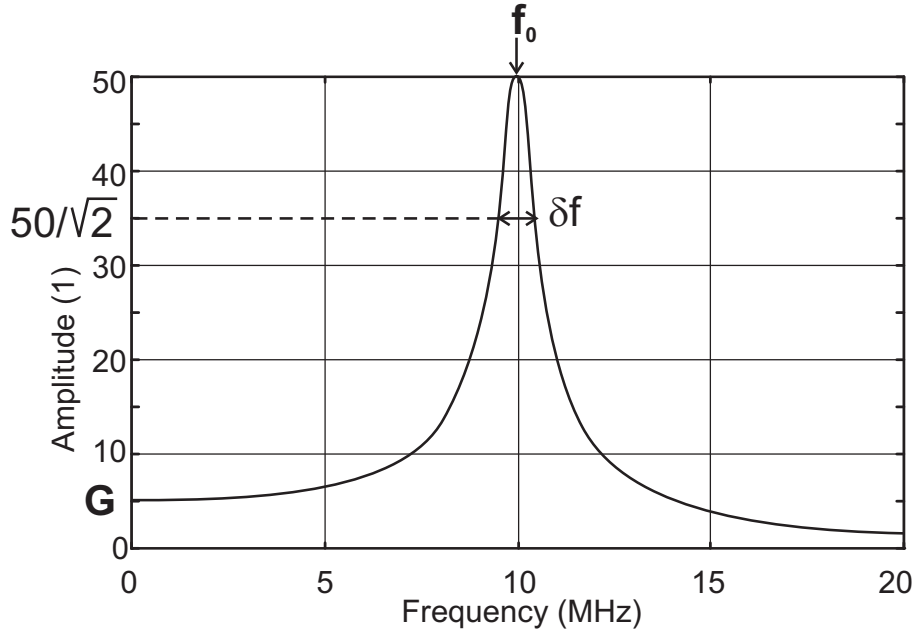
We have plotted in the figure the transfer function of a second order system, and we directly get : , , and  $\delta f \approx 1$  MHz. From these last datas we obtain:

- the static gain,  $G \approx 5$ ,
- the natural frequency  $f_n \approx 9.8$  MHz, by observing that for the visibly sharp resonance we have,  $f_n \approx f_0$ ,
- $\zeta = 1/2Q \approx 0.05$ , by observing that  $Q = \omega_0/\delta\omega = f_0/\delta f \approx 10$ .

The differential equation governing this system may thus be written as:

$$y \approx 5x - 10^{-8}dy/dt - 10^{-14}d^2y/dt^2$$

Alternatively the system could be represented by a RLC circuit with elements connected in series, and, from Table 2.4  $C = G$ ,  $L \approx 1/G\omega_0^2$ ,  $R \approx 2\zeta/G\omega_0$ .



### 2.5.4.3 Quality factor

There is another way to look at the resonance by using the concept of *quality factor*. We have seen that high damping factor are associated with small resonance peak or even no peak at all, and conversely low damping means a high resonance peak. The damping of a system represents its loss to the environment, usually as heat, and a system that has a large loss is understandably a low quality system. Thus we also quantify the sharpness of the resonance using the quality factor, defined as the maximum of the normalized frequency response amplitude, that is the frequency response when  $\omega = \omega_0$ :

$$Q = \frac{1}{2\zeta\sqrt{1-\zeta^2}} \approx \frac{1}{2\zeta}$$

However this definition of the quality factor does not give much more information than the damping ratio, and is not very easy to measure practically as far from the resonance the amplitude may be too small and lost in the system noise. A more practical definition, valid for  $Q \gg 1$ , is given by:

$$Q \approx \frac{\omega_n}{\omega_2 - \omega_1} \approx \frac{\omega_0}{\omega_2 - \omega_1}$$

where  $\omega_1$  and  $\omega_2$  are the frequency where the transfer function is  $\approx \frac{1}{\sqrt{2}} \frac{G}{2\zeta}$ , that is, the maximum of the transfer function divided by  $\sqrt{2}$ . As signal energy is proportional to the square of its amplitude, these two points correspond to frequencies where the energy in the system has been reduced to half of the maximum energy stored at resonance.

This definition is quite easy to apply and is used to define a quality factor and a damping ratio, even when the system is not purely a second-order system.

## 2.5.5 Effect of frequency on system response

A good insight in the frequency limitation of a system is obtained with the Bode plot. Actually the change of the amplitude and phase of the transfer function with the frequency of the input signal means that the characteristics of the system will vary with the frequency of the input signal! Meaning, for example, that a micro-sensor measuring two measurands with the same magnitude could return different informations because the two measurands have different frequencies. This effect is often difficult to compensate, thus usually systems are rated for a maximum frequency above which the manufacturer does not warranty their characteristics anymore.

Let's have a look at what it means for our typical systems, looking separately at the limitations imposed by the amplitude and the phase change.

### 2.5.5.1 Gain distortion

The change in the amplitude of the transfer function place a limit to the range of frequency where a system will keep its accuracy. To understand the problem

we will imagine that we have a photodiode, modeled here as a first order system with cut-off frequency  $f_c$  ( $f_c = 2\pi/\tau$ ), and that we need to compare two luminous signals modulated at different frequencies. When one of the modulation frequency is much smaller than  $f_c$  the sensitivity of the sensor is about  $G$ , but if the frequency is much larger than  $f_c$  the sensitivity of the sensor approaches quickly 0, and the sensor won't 'see' the modulation at all! In this case a comparison between the two measurements is barely possible, because the sensor will have *different output signal* for signal that originally have the *same amplitude*, simply because they are at different frequencies! Figure 2.15 illustrates this effect by using an input whose spectrum has two frequency components, one at  $0.5f_c$  the other at  $7f_c$ . The two frequency component have the same amplitude and the system transfer function spectrum is representative of a first order system. In the output signal, the two components don't have the same amplitude anymore, because the system did not show the same gain at all frequencies. This results in wrong measurement, making signal comparison almost impossible.

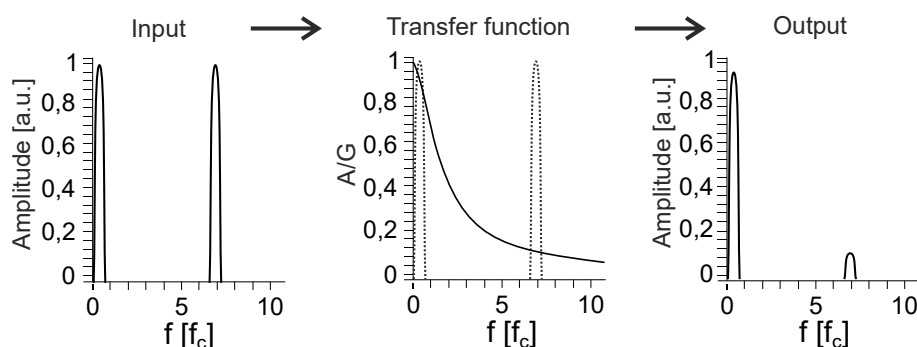


Figure 2.15: Measurement error induced by the limited bandwidth of a sensor.

For a first order system, if we want to keep the change in gain smaller than 1% for all frequencies of operation, we need to have  $A(\omega) > 0.99A(0)$ , where  $A(\omega)$  is the amplitude of the transfer function  $H(j\omega)$ . Using the expression for  $A(\omega)$  given in eq. 2.10 we find that this condition is fulfilled if  $f < 0.14f_c$ : we thus need to keep the operating frequency below 15% of the cut-off frequency to maintain accuracy within 1%.

If the system is represented by a second order model, we may also try to find an estimation of the maximum operating frequency to keep the gain error (or distortion) within 1%. In Figure 2.16 we zoom in the previous Bode plot for frequency smaller than the resonance frequency to obtain the information. We see that if we want to keep the error within 1%, it is necessary to keep the frequency of the input below  $0.1f_n$  if the damping is very small (i.e.,  $\zeta \approx 0$ ). When the damping is large ( $\zeta \geq 1$ ) the limitations becomes quickly even more drastic, allowing, for example, a maximum operating frequency of only 2% of  $f_n$  for  $\zeta = 5$ . In the other hand, we see the interest of having a damping ratio around 0.7: it increases significantly the bandwidth of the system. In this case the bandwidth (assuming

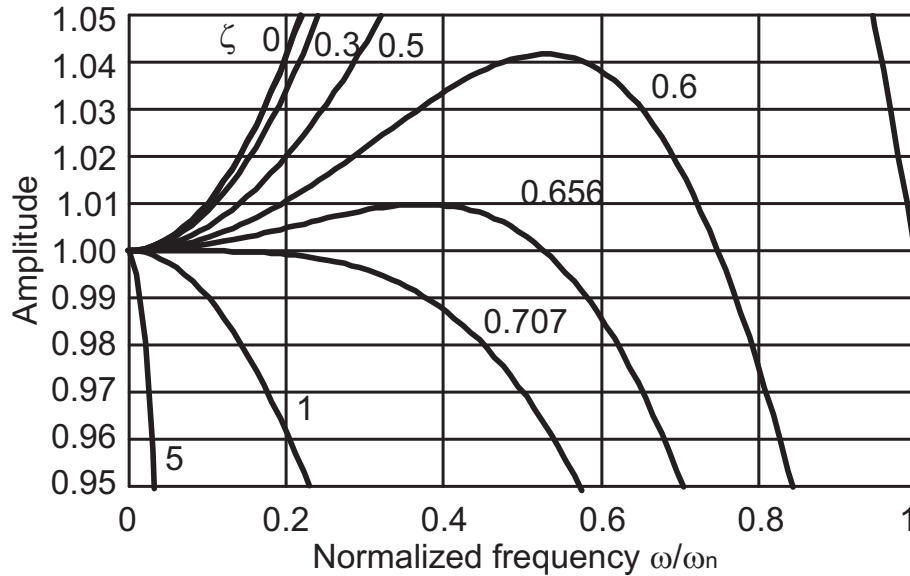


Figure 2.16: Amplitude plot for a 2nd order transfer function with different damping ratio  $\zeta = [0, 0.3, 0.5, 0.6, 0.656, 1/\sqrt{2}, 1, 5]$ .

again that an error smaller than 1% can be tolerated) even exceed 50% of  $f_n$  if  $\zeta = 0.656$ .

System response will always benefit from a damping ratio around 0.7, even if this may be difficult to implement in practice. It should be noted, that active structures may be used to circumvent this problem : for example using a feedback loop and an actuator inside a micro-accelerometer may help to increase the operating frequency up to the resonance frequency, even for systems naturally badly damped. Of course, in that case the complexity of the system increases substantially.

### 2.5.5.2 Phase distortion

In the previous section we have only considered the effect of the frequency on the amplitude of the transfer function (i.e. the gain or sensitivity), however the phase is also varying with the frequency and we may ask ourselves what is the effect of this variation on the signal. This issue will affect first-order and second-order systems but it will be more pronounced with second-order system because their bandwidth (determined on amplitude consideration as above) may be larger than for first order system, and because the maximum phase change is two times larger.

Actually if the input signal is a ‘purely’ sinusoidal function of time, the change in phase is not important because usually the phase of the output signal is not a relevant factor. Thus even if the output phase varies with the frequency, the amplitude won’t be affected causing no problem with systems where the information is contained in the amplitude of the output signal.

However a problem arises when the signal has a complex time dependence. To describe the reasoning we will consider a periodic input signal and a second-order sensor, but the result holds for all kind of signal and systems as well. As the signal is periodic, we may use the Fourier's decomposition to represent the input signal as a sum of sinusoids, whose amplitude is obtained from the signal spectrum. For example, as shown in Figure 2.17, the signal is decomposed as a fundamental and its first harmonic at a frequency twice the fundamental frequency. For the sensor we take  $\zeta \approx 0.05$  (small damping) and we consider that the signal fundamental frequency is about  $\omega_f = 0.5\omega_n$ .

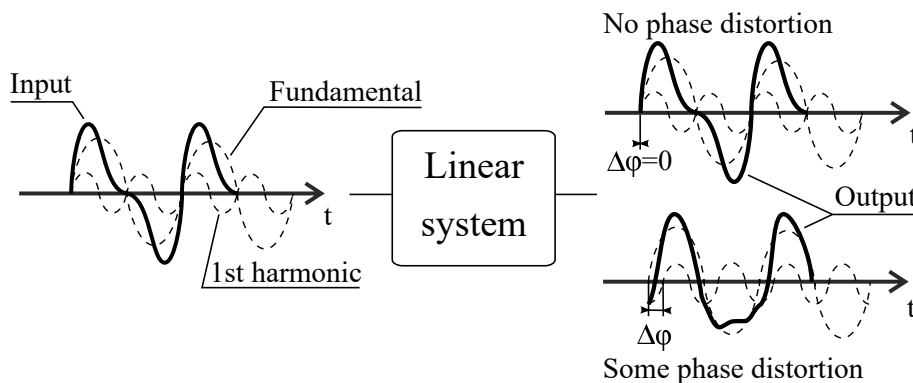


Figure 2.17: Phase distortion effect.

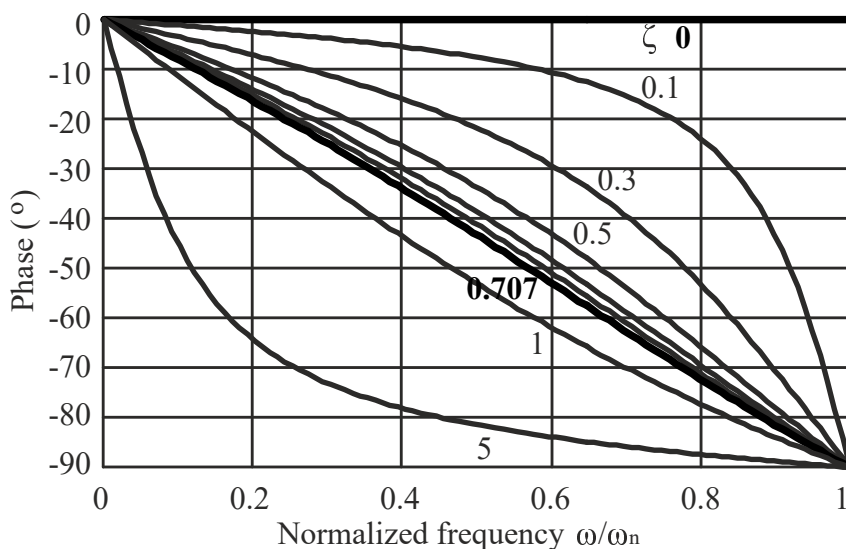


Figure 2.18: Phase plot for a 2nd order transfer function with different damping ratio  $\zeta = [0, 0.1, 0.3, 0.5, 0.656, 1/\sqrt{2}, 1, 5]$ .

From the plot of Figure 2.18 we see that the sinusoidal component at the fundamental frequency will have essentially no phase shift, while the first harmonic at  $\omega_1 = \omega_n$  will experience a  $90^\circ$  shift. Neglecting the effect of amplitude distortion,

we plot the fundamental and the first harmonic on a time scale as in Figure 2.17 by remembering that  $\phi = \omega t$ . As the system is linear, we can use the so called superposition theorem and the output is given by the sum of the two individual responses given by each of the two sinusoidal components of the input signal. We see that the output corresponding to the first harmonic is shifted by a quarter of the period of the fundamental resulting in a large distortion of the output signal, that doesn't look at all similar to the original signal!

For alleviating this problem in a system there are two possibilities, either all the sinusoidal components of the input signal should experience no phase shift (no-shift), or they should experience the same shift on the time scale (linear-shift). The same shift on the time scale means that the phase shift has to be linearly proportional to the frequency as we have

$$\begin{aligned}\phi &= \omega t \\ \Rightarrow \Delta\phi &= \omega\Delta t \Rightarrow \Delta t = \Delta\phi/\omega \\ \text{If } \Delta t &= \text{const} \\ \Rightarrow \Delta\phi/\omega &= \text{const} \Rightarrow \Delta\phi = \text{const } \omega.\end{aligned}$$

For the discussion we will only consider the component that have a frequency smaller than the natural frequency ( $\omega \leq \omega_n$ ), because the other components of the spectrum would have a negligible amplitude as we have seen in the previous section (the amplitude of the transfer function quickly drops after  $\omega_n$ ).

From the Figure 2.18, the no-shift condition is fulfilled for any frequency when the damping ratio is very small ( $\zeta \ll 1$ ) or when the bandwidth is severely limited  $\omega \ll \omega_n$ .

The linear-shift condition is approximately fulfilled when the damping ratio is about  $\zeta \approx 1/\sqrt{2}$ . Actually looking at Figure 2.18 we see that we have  $\Delta\phi \approx -\frac{\pi}{2} \frac{\omega}{\omega_n}$ , thus  $\Delta\phi$  changes linearly with  $\omega$ , just what we want.

In conclusion, we see that for a second order system it is very desirable to have a damping ratio around 0.7, because it will simultaneously increase the bandwidth of the system and decrease its phase distortion. In general, though, the effect of phase distortion is less watched during design because high frequency signals suffering a large phase distortion will anyway see their amplitude heavily attenuated. Moreover, one may notice by comparing the curves in Figure 2.16 and 2.18, that the effect of a small change of  $\zeta$  around  $\zeta = 0.7$  has much more impact on the amplitude than on the phase. We may thus expect that it will be more difficult to control exactly the gain than the phase distortion if we choose to place ourself near  $\zeta = 0.7$ .

The simplest approach to control the effect of frequency on system response is to severely limit the operation frequency with respect to the resonant or cut-off frequency – or equivalently to push these frequencies towards the higher end of the spectrum.



## 2.6 Advanced sub-systems modeling

### 2.6.1 Multi-domain circuit elements

MEMS are very often multi-physics system working in more than one energy domain. Actually, many MEMS are used to transform energy in one energy domain to another one and are called *transducers*. A typical example will be a motor, that is a system that convert electrical energy to mechanical energy. But a microphone would also be such a device, as it converts acoustic/mechanical energy to electrical energy.

If this is true for a system it can also be found for simple elements in sub-systems, that are used to convert energy between domains. The one port circuit elements we have encountered earlier are intrinsically single domain, however the two-ports elements (Fig. 2.10) can be used for describing elements that work in two domains. One port will be used for connecting to other elements in a certain energy domain (e.g., mechanical domain) while the other will be used to connect to another domain (e.g. electrical domain).

Actually in its simplest form, a linear reciprocating<sup>13</sup> multi-domain element is a linear quadrupole. As such they can be represented with different impedances in conjunction with a transformer, where the transforming ratio  $1 : n$  actually has the dimension suitable to link the effort and flow variables in both domains (Fig. 2.19). When we work with transformer, we repeatedly make use of the

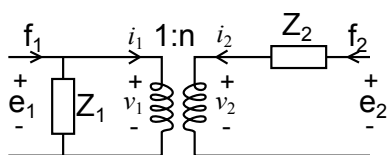


Figure 2.19: Circuit model of a reciprocating linear multi-domain system.

relationships existing between the two effort variables at its two ports:

$$v_2 = nv_1$$

accordingly, because it is an energy conserving device and  $v_1 i_1 = v_2 i_2$ , the two flow variables are linked using

$$i_2 = \frac{1}{n} i_1$$

Additionally the transformer has impedance transforming properties. Actually, an impedance  $Z_2$  connected between the two terminals of the secondary of the transformer, is seen from the primary side, as an impedance, connected in parallel with the primary port, of value:

$$Z_1^{\text{eq}} = \frac{v_1}{i_1} = \frac{\frac{1}{n} v_2}{n i_2} = \frac{1}{n^2} \frac{v_2}{i_2} = \frac{1}{n^2} Z_2$$

<sup>13</sup>that is the transfer between energy domain is equal both way, which happens in any energy-conserving (lossless) transducer

For example a capacitance  $C$  of impedance  $Z_C = 1/(jC\omega)$  connected to the secondary, will be equivalent to an impedance  $Z_C^{eq} = 1/(jn^2C\omega)$  seen from the primary, that is it could be represented by a capacitor of value  $n^2C$ .

---

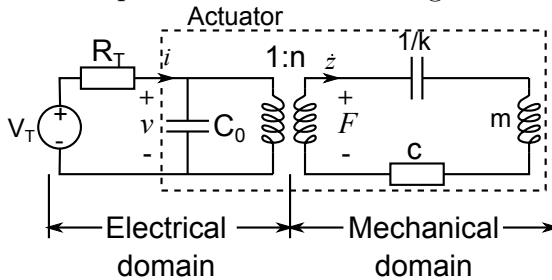
**Example 2.10** Circuit modeling of a piezoelectric actuator.

---

THE PIEZOELECTRIC actuator is a linear actuator based on the converse piezoelectric effect, where, in some materials, the application of an electric field causes the apparition of strain. This is the typical example of a (quasi) lossless conversion mechanism between the electrical and the mechanical domain. It is actually a reciprocating linear process, that can be described by the linear formalism we developed earlier for multi-domain elements.

A voltage is applied on the electrodes on both side of the material plate that induces strain and stress deforming the plate until it is balanced by a resisting elastic stress. At the same time the deformation changes the polarization of the material (direct piezoelectric effect), in turn modifying the charge on the electrodes. From its electric behavior the actuator is a capacitor, whereas in the mechanical domain, it is a classical second order system.

Accordingly the following circuit model can be used for studying the dynamic of the complete actuator including the excitation by a voltage source:



If we consider the effort variables on both sides of transformer we have  $e_2 = ne_1$ , the dimension of  $n$  should link the force  $F$  in the secondary to the voltage  $v$  in the primary, and thus should be Newton/Volt. Another way to look at it using the flow variables and we have  $f_2 = \frac{1}{n}f_1$ . Then  $n$  should link the current  $i$  to the velocity of the actuator  $\dot{z}$  and is thus expressed as Ampere/Meter per second, or if we drop the time, as Coulomb/Meter.

We can use dimensional analysis to verify that these two units are indeed the same:  $[NV^{-1}] \equiv [N(J/C)^{-1}] \equiv [NCJ^{-1}] \equiv [NC(Nm)^{-1}] \equiv [Cm^{-1}]$ .

Computing the different expression of the lumped elements appearing in the model will be done in Example 4.5.3.

---

## 2.6.2 Non-linear sub-system dynamics

The methods we have described in the previous sections are mostly designed for linear systems, and although we have suggested that the equations governing physics are linear, they often yield non-linear solutions. For example, if we consider

the Coulomb's force between two electric charges,  $q_1$  and  $q_2$ , we have:

$$\vec{F}_{1/2} = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r_{12}^2} \hat{r}_{12}$$

If we consider the charge  $q_1$  (thus we are in the electrical domain), the force is linear, and the effect of additional charges would be easily obtained with the superposition theorem... however if we are in a system where the distance  $r_{12}$  can vary (we place ourselves in the mechanical domain), as would happen in an electrostatic actuator, then the law becomes non-linear!

In that case, block or circuit model (with non-linear elements) can still be used for modeling - but Laplace's transform is no more useful for studying the system dynamics. In that case one would have to use state equations, that is write – and solve numerically – a set of first order differential equations describing the system<sup>14</sup>.

Actually, one should write one equation per independent state variable  $u_i$ , that is, per independent energy storage elements (i.e., per capacitor and inductor in the analogue circuit), and the equations will give the derivative of each state variables  $\dot{u}_i$  as a function of the state variable (but not their derivatives) and the inputs of the system  $x_i$ .

$$\begin{cases} \dot{u}_1 = f_1(u_1, u_2, \dots, x_1, x_2, \dots) \\ \dot{u}_2 = f_2(u_1, u_2, \dots, x_1, x_2, \dots) \\ \dot{u}_3 = f_3(u_1, u_2, \dots, x_1, x_2, \dots) \\ \vdots \end{cases} \quad (2.16)$$

As the system is non-linear at least one of the  $f_i$  will be non-linear and the solution will require in most cases numerical solution to find the system dynamics. These equation can be directly integrated numerically or alternatively can be placed in the MATLAB Simulink<sup>®</sup> environment. One may note that the steady state of the system is obtained when the right hand side of each equation is equal to 0 (i.e. all  $\dot{u}_i = 0$ ). The output  $y$  of the system can be used as one of the state variables but it is not always possible, and in this case it will be derived from the state variables  $y = g(u_1, u_2, \dots)$ .

---

<sup>14</sup>Clearly state equations can be used to represent linear system too, but the Laplace transform or the complex impedance are simpler to use

---

**Example 2.11** Dynamic model of a non-linear electrostatic actuator.

---

WE CONSIDER the electrostatic actuator on the right where the voltage  $V$  controls the position of the upper electrodes  $x$ .

We can model the dynamic of this structure by writing Newton's second law on the mobile plate, where  $m$  will be the rotor mass and  $c$  represent the air damping, while the first term is the electrostatic force obtained in Sect. 4.5.2:

$$\epsilon_0 \frac{A}{2x^2} V^2 - k(x - g) - c\dot{x} = m\ddot{x}$$

The electrical circuit in the other hand is rather simple and we write:

$$-V + RI + \frac{1}{C} \int I dt = 0$$

we get the integral disappear by using the charge  $Q = \int I dt$  instead of the current, and get:

$$-V + R\dot{Q} + \frac{Q}{C} = 0$$

where the capacitance  $C = \epsilon_0 \frac{A}{x}$  in the parallel plate approximation, thus finally giving:

$$-V + R\dot{Q} + \frac{xQ}{\epsilon_0 A} = 0$$

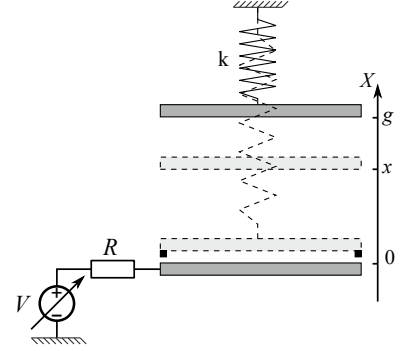
In this circuit there are 3 energy storage elements: two in the mechanical domain (the inductor of the mass and the capacitor of the spring), and one in the electrical domain, the capacitor formed by the electrodes. We will need 3 independent state variables that should yield only first order differential equations, that is, we need  $u_1 = \dot{x}$  and  $u_2 = x$  to be able to bring the mechanical equation to a proper format. The last variable will be  $u_3 = Q$  as it appears as  $\dot{Q}$  in the electrical equation and we can finally write the state equations :

$$\begin{cases} \dot{u}_1 = \frac{1}{m} \left( \epsilon_0 \frac{A}{2u_2^2} V^2 - k(u_2 - g) - cu_1 \right) \\ \dot{u}_2 = u_1 \\ \dot{u}_3 = \frac{1}{R} \left( V - \frac{u_2 u_3}{\epsilon_0 A} \right) \end{cases}$$

We note here that the output of the system, the position of the actuator  $x$ , is directly obtained from the state variables as we have :

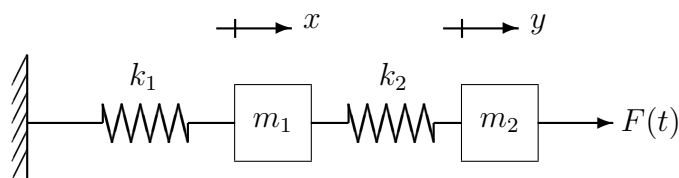
$$x = u_2$$


---



## Problems

1. The Mad Hatter wants to serve Alice tea, but the tea is waaaaay too hot. From basic principles, will it be faster (White Rabbit is waiting !) to let the tea cool in the teapot before pouring it in the cups or to empty the teapot in the cups first and let the tea cool down there? How much faster ? (The Hatter's teapot makes  $2\pi$  cups - but we would count it as 6 cups and consider every container as spherical)
2. Starting from the expression of the frequency response of a second order system and using the definition of the Q-factor in the frequency domain (the difference between the frequencies where the energy in the system has been reduced to half of the maximum energy stored at resonance divided by the resonance frequency.), establish the exact expression for Q and simplify it in the case where  $\zeta \ll 1$ .
3. Find the circuit representation of the lumped mechanical system in the Figure.





# Chapter 3

## How MEMS are made

### 3.1 Overview of MEMS fabrication process

Micro-fabrication is the set of technologies used to manufacture structures with micrometric features. This task can unfortunately not rely on the traditional fabrication techniques such as milling, drilling, turning, forging and casting because of the scale. The fabrication techniques had thus to come from another source. As MEMS devices have about the same feature size as integrated circuits, MEMS fabrication technology quickly took inspiration from microelectronics. Techniques like photolithography, thin film deposition by chemical vapor deposition (CVD) or physical vapor deposition (PVD), thin film growth by oxidation and epitaxy, doping by ion implantation or diffusion, wet etching, dry etching, etc have all been adopted by the MEMS technologists. Standard book on microelectronics describe in details these techniques but, as MEMS and IC fabrication goals are different, some of these techniques have evolved as they were applied to MEMS and we will detail here their new capabilities. Moreover, MEMS has spurred many unique fabrication techniques that we will also describe in our panorama of MEMS fabrication introducing bulk micromachining, surface micromachining, LIGA, etc [21].

In general, MEMS fabrication tries to use batch process to benefit from the same economy of scale that is so successful in reducing the cost of ICs. As such, a typical fabrication process starts with a wafer (silicon, polymer, glass...) that may play an active role in the final device or may only be a substrate on which the MEMS is built. This wafer is processed with a succession of processes (Table 3.1) that add, modify or remove materials along precise patterns. The patterns (or the layout) is decided by the designer depending on the desired function but, for most materials, it is difficult to directly deposit or modify them locally. In fact there are few processes equivalent to turning or milling in micromachining. Focused Ion Beam (FIB), where a beam of high energy ion can be scanned to remove most materials and deposit some, can perform even down to nanoscale but the sequential processing approach it requires (as opposed to batch processing) is not cost

Additive process	Modifying process	Subtractive process
Evaporation	Oxydation	Wet etching
Sputtering	Doping	Dry etching
CVD	Annealing	Sacrificial etching
Spin-coating	UV exposure	Development
...	...	...

Table 3.1: Process classification.

effective for production. Thus the problem of patterning a material is generally split in two distinct steps: first, deposition and patterning of a surrogate layer that can be easily modified locally and then transfer of the pattern to the material of interest.

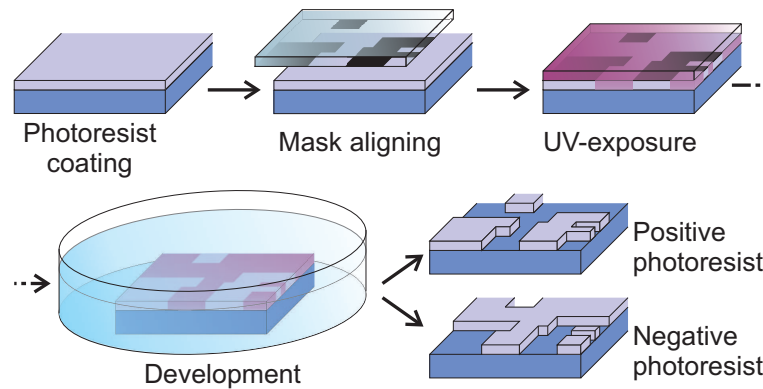


Figure 3.1: Photo-patterning in positive and negative photoresist.

In the most common process called photo-patterning, the surrogate layer used is a special solution (called a photoresist) which contains a polymer sensitive to UV-photon action (Figure 3.1). The liquid photoresist is first coated on the substrate as a thin-film and the solvent is then evaporated by baking on a hotplate leaving a solid layer of polymer with very uniform thickness. The substrate is then brought to the mask aligner tool, where the patterning process will take place. The photoresist film is exposed to UV radiation through a mask which has been precisely aligned with the substrate. The mask has clear and opaque regions according to the desired pattern, the clear regions allowing the photoresist to be exposed to UV radiation and modifying it locally. The exposure creates a latent image of the mask feature in the surrogate layer. The contrast of this image may be enhanced by heat, which accelerates the chemical reaction initiated by the UV-exposure. To finish the process this latent image is revealed in a special chemical solution, the



developer. Actually the exposure changes the solubility of the photoresist in the developer and the exact change of solubility depends on the type of photoresist used originally: for so-called positive photoresist the exposed region becomes more soluble in the developer, while for negative photoresist the reverse happens and the exposed region becomes insoluble. After development, the surrogate layer patterned over the whole surface of the wafer can be used for pattern transfer.

They are actually two main techniques that can be used to transfer the pattern:

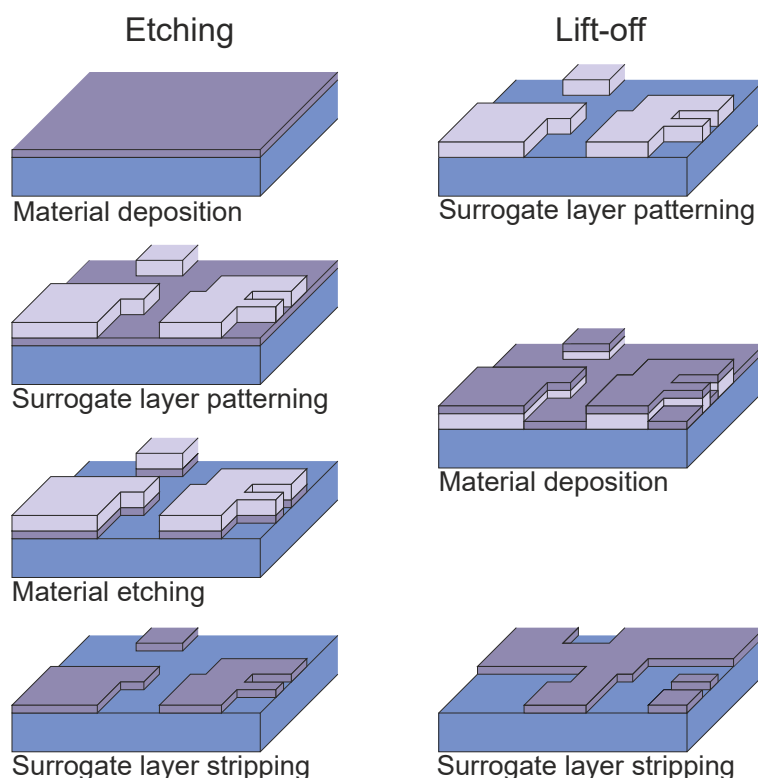


Figure 3.2: Pattern transfer by etching (left) and lift-off (right).

etching and lift-off (Figure 3.2). With etching the patterned layer allows protecting locally the underlying material. The material in the unprotected regions is then attacked physically or chemically before we finally remove the protective layer. For lift-off, the material is deposited on top of the patterned layer. Complete dissolution of this layer (used actually as a sacrificial layer) simultaneously washes away the deposited material it was supporting, leaving the deposited material only in the open regions of the pattern.

Combination of photo-patterning and etching is known as photolithography and is nowadays the most common techniques for micro-fabrication, lying at the roots of the IC revolution. An important step in this process is the fabrication of the photolithographic mask, which can clearly not itself use photolithography with a mask! Its fabrication is actually based on etching of a chromium film deposited on a transparent substrate. The pattern is obtained inside a pattern generator

system where the photoresist film is exposed point by point by a beam (electron or laser) which is slowly scanned over the whole mask surface while its intensity is modulated by a computer control system. This process reaches a very high resolution (electron beam can be focused in spot as small as 10nm) but it is slow. This remains an acceptable trade-off for mask production, as the photolithographic step requires to produce a unique mask which is then repeatedly used to expose 1000's of wafers.

Recently we have seen the emergence of new patterning techniques that try to reduce the cost attached to photo-patterning at small scale where there is a need to use deep-UV source with complex optics using immersion lens and systems under vacuum. The most promising ones are based on imprinting, where the desired pattern on a stamp is pressed against a protective resin film (hot-embossing, nano-imprint lithography, UV-NIL...). The resulting patterned layer can then be used with lithography or lift-off for pattern transfer.

This succession of patterning/pattern transfer is repeated a certain number of times on different layers. Usually only a few layers are patterned and stacked together, but the production of more complex MEMS may use a stack of about 10 different layers - still a far cry from the IC process that could necessitate more than 30 different patterning steps.

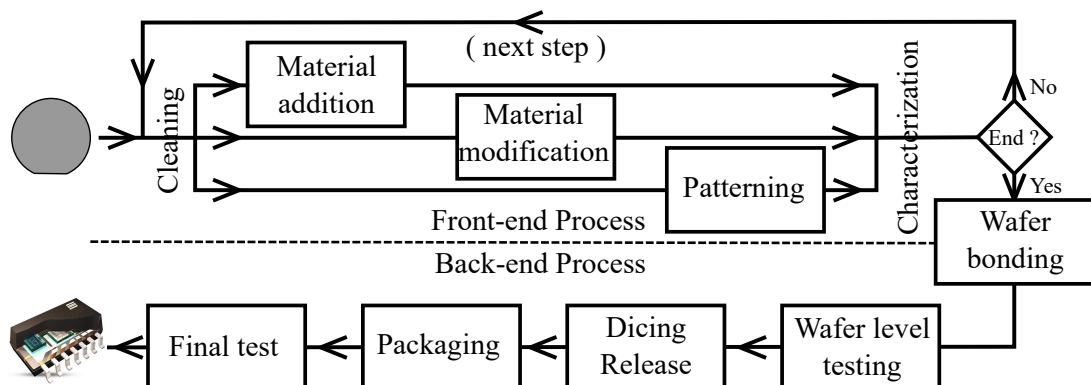


Figure 3.3: General view of MEMS production process.

MEMS production (Figure 3.3) does not end with wafer fabrication, the so-called front-end process, and there are, like in electronics, back-end processes... with a twist. For example, when ‘mechanical’ parts (mobile elements, channels, cavities...) exist in the device, the processed wafer passes through a special step called ‘Release’ to free these parts. This step may happen before or after the wafer is diced in individual chips, just before the more traditional assembly, packaging and final tests.

There is a fundamental aspect of all the micro/nanofabrication operations that has not been mentioned yet: it needs to be conducted in a dust free environment. In fact, when dust particles are bigger than the things you’re fabricating, it is clear that dust needs to be kept at bay! However many sources of dust exist:

combustion (smoke, gas...), living species (pollen, virus, bacteria, mites feces...), machines (lint) and, most importantly, humans (skin cells, face powder, tobacco smoke, hair...). Actually a person who is motionless will generate about 100,000 particles  $0.3 \mu\text{m}$  and larger per minute - and that number climbs to 5 millions if that person walks slowly ! Thus there is no possibility to have an ideal dust free place, and what we need is a way to continuously remove the generated dust particles out of the wafer way. This is obtained by working in clean-rooms, specially

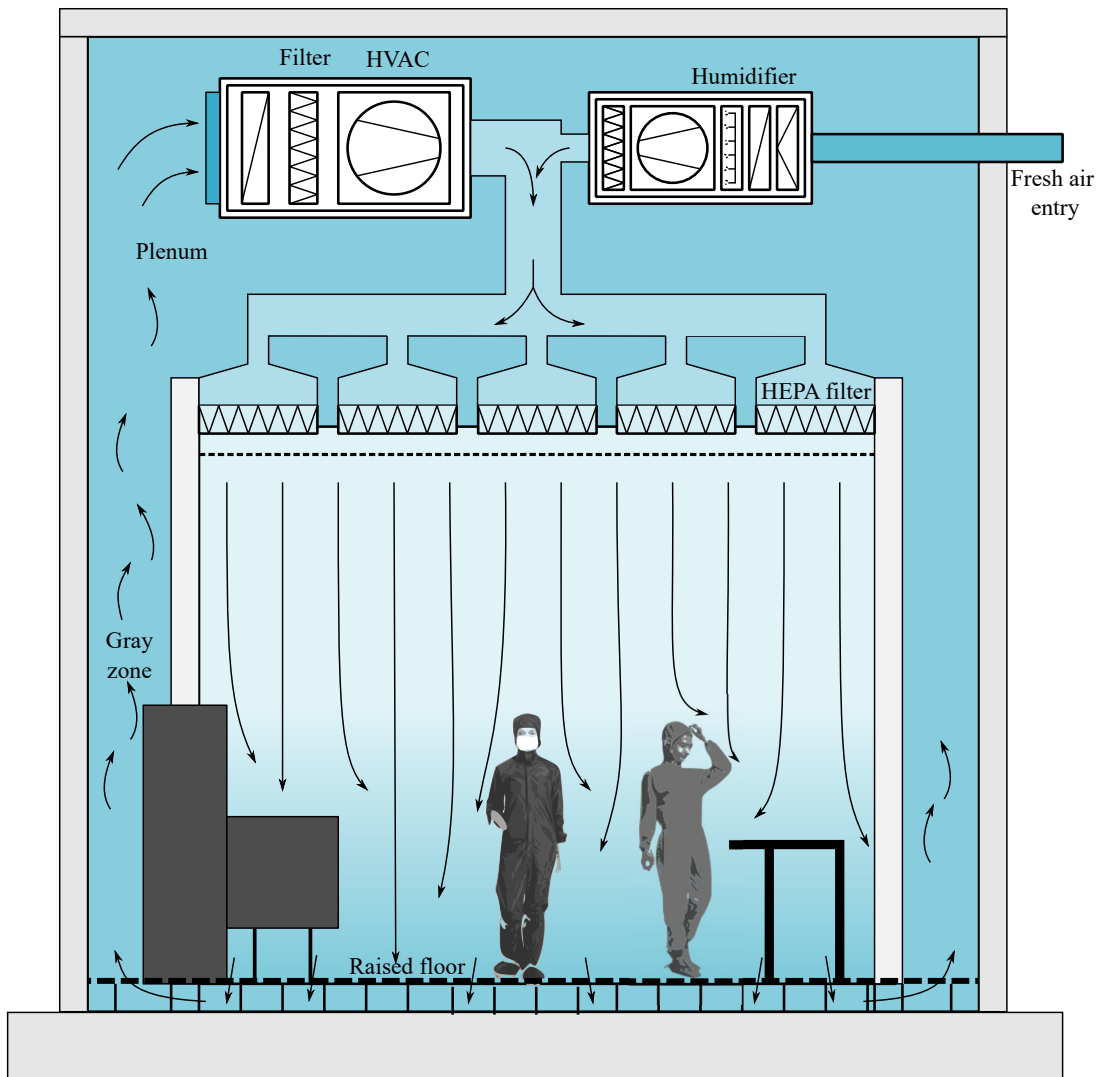


Figure 3.4: Principle of operation of a cleanroom.

built workshop where the level of particle is kept to a very low level, as shown in Figure 3.4. The principle is simple, we force laminar air flow from the ceiling to the floor to pull down the dust particles in air towards the floor as fast as possible. Then the air is recirculated after passing through the floor (or at the bottom of the wall) and the gray zone that is used to holds part of the machine where the operator

don't need access. The dirty air is then cooled (it carries all the calories generated by machine in the room) and re-injected in the room through high quality filter (HEPA filter will remove 99.99% of particles 0.3  $\mu\text{m}$  or larger and more than 95% of particles above 10 nm) to create the laminar flow. Moreover, a small quantity of fresh air is continuously being injected after treatment, both for maintaining a good air quality for the users and also for "refilling" the room as the chemical benches are continuously evacuating air from the room to avoid exposing users to chemical vapour. But the clean-room function does not stop at dust control, and the clean-room infrastructure actually controls 4 main characteristics:

- Force vertical laminar flow of clean air for pushing particles to the floor (gravity alone will take 48 min to pull a 1  $\mu\text{m}$  particle 1 m down). The HVAC (Heating, Ventilation and Air Conditioning) machinery replaces the volume of air 60-300 times/h, depending on the level of cleanliness required.
- Maintain fixed temperature (around 21°C) with HVAC for keeping process (chemistry) repeatable
- Maintain a fixed humidity (relative humidity between 20% and 50% at  $\pm 5\%$ ) for keeping process (water surface adsorption) repeatable. This parameter can not be too low to avoid electrostatic charging and it is linked with temperature (RH varies with temperature and water vapour pressure). Maintaining humidity constant consumes the most energy, as air is cooled to dry it before it is heated and humidified to the right level.
- Impose an overpressure (+10 to +30Pa) to keep dust outside at the entrance and anywhere there is a leak in the clean-room envelope.

The cleanliness of the clean-room is indicated by its ISO class, ISO X, where X is the log of the number of particle 0.1  $\mu\text{m}$  or bigger per  $\text{m}^3$  as shown in Table 3.2. We may note that the old FED standard (enunciated as "Class C"), based on imperial units (C is the number of 0.5  $\mu\text{m}$  particle or bigger per  $\text{ft}^3$ ), is still regularly used. The smaller the class, the higher the cost, as for example, the air replacement rate becomes larger and larger, requiring more and more HEPA filters and larger machinery. The design of clean-room will thus involve minimizing the area of low ISO numbers and reserve it where it is needed the most. In fact the room where patterning or wafer bonding are performed, are the most sensitive and in academic environment are set at ISO5 or better, whereas room for materials deposition may be at ISO6 or 7.

In industrial clean-rooms using large wafers, the manual operation is almost completely removed and actually the wafer are moved between automated equipment using robotic handler and enclosed environment called FOUP (Front Opening Unified Pod). In that case, the requirement in the clean-room itself is less stringent, as the wafer never sees the air there, but only the confined space of machine chambers and the FOUPs, that can be both more easily maintained at very high level of cleanliness (ISO3 equivalent).

Class	$\geq 0.1 \mu\text{m}$ particles/ $\text{m}^3$	$\geq 0.5 \mu\text{m}$ particles/ $\text{ft}^3$	Old FED STD 209E
ISO1	10		
ISO2	100		
ISO3	1000	1	Class 1
ISO4	$10^4$	10	Class 10
ISO5	$10^5$	100	Class 100
ISO6	$10^6$	1000	Class 1000
ISO7	$10^7$	$10^4$	Class 10,000
ISO8	$10^8$	$10^5$	Class 100,000
ISO9	$10^9$	$10^6$	Room air

Table 3.2: Cleanroom cleanliness standards.

To imagine the full complexity of a clean-room, we should not forget that the clean-room facilities should also provide all the fluids for running the machines and the process. At minimum this includes:

- Cooling water for machines and for the air plant (dehumidifier)
- Clean and deionized water for process
- Process gases for deposition and etching of materials
- Inert gases for operating machines, purging chambers, ...
- Compressed air for operating machines

In fact, even without any process run, a clean-room is a very costly environment and should be appreciated by all its users at its full value!

## 3.2 MEMS materials

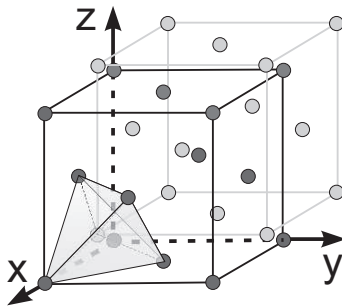
### 3.2.1 Crystalline, polycrystalline and amorphous materials

Broadly speaking, the MEMS materials can be split in three classes depending on how orderly the atoms are arranged in space: at one extreme, the crystalline materials, where order prevails; at the other end, the amorphous materials, where orientation varies wildly between neighbouring atoms; and in between, the polycrystalline materials, where order is conserved only on a short scale, called a grain, while on a larger scale it is made of arrangement of differently oriented grains.

A single crystal presents the highest order, as the atoms are periodically arranged in space in a precise manner following one of the lattice allowed by thermodynamic and geometry, that is, one of the 14 Bravais' lattice. The crystal is then built by the repetition of this elementary lattice in all three directions. In this case the material properties of the crystal are highly reproducible but they will generally depend on the direction within the crystal, and the material is said to be anisotropic.

In the case of polycrystalline films, the material does not crystallize in a continuous film, but in small clusters of crystal (called grains), each grain having a different orientation than its neighbour. In general the grain size range from about 10 nm to a few  $\mu\text{m}$ . The grains may not be completely randomly oriented and some direction may be favored depending on the material elaboration process, resulting in highly varying material properties for different process condition. If the distribution of grain orientation is known, a good approximation of the properties of the material can be obtained by using the weighted average of the single crystal properties along different directions.

Finally, in amorphous films, the material grows in a disordered manner, with clusters of crystal being of a few atoms only. In this case, the material properties are not the same as those present in single crystal or in polycrystalline films, and usually present inferior characteristics: lower strength, lower conductivity... Interestingly, the properties of amorphous material are normally more stable with their elaboration process parameters and they are also independent of the direction : amorphous materials are intrinsically isotropic.



In a crystal it is possible to identify the position of the atoms using a system of coordinate whose axes are placed along the edge of the lattice. In a cubic crystallographic system (as Si, AsGa...), the coordinate system is Cartesian with the usual  $x, y, z$  axes. Moreover the coordinate along these axes are usually expressed in lattice unit, that is, the atom at lattice origin has the coordinates  $0, 0, 0$ , while the one at the furthest position is located at  $1, 1, 1$ . The

lattice describes the arrangement of the unit cell, which is not always composed of a single atom. For example in GaAs one unit cell is made of one atom of As and one atom of Ga. Similarly, for Si crystal, the elementary cell has two Si atoms, one placed in  $0, 0, 0$  and one in  $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}$  which are arranged following a face-centered cubic (fcc) configuration. In this case (2 atoms in the elementary cell), we can also see the atom arrangement as two fcc lattices offset by a quarter of the diagonal. On this configuration, we have highlighted the tetrahedron configuration that shows that each Si atom (grey) has 4 neighbours (black), sharing one electron with each as silicon is a tetravalent material. Many semiconductor materials - usually tetravalent - will share this fcc arrangement, also called the diamond lattice.

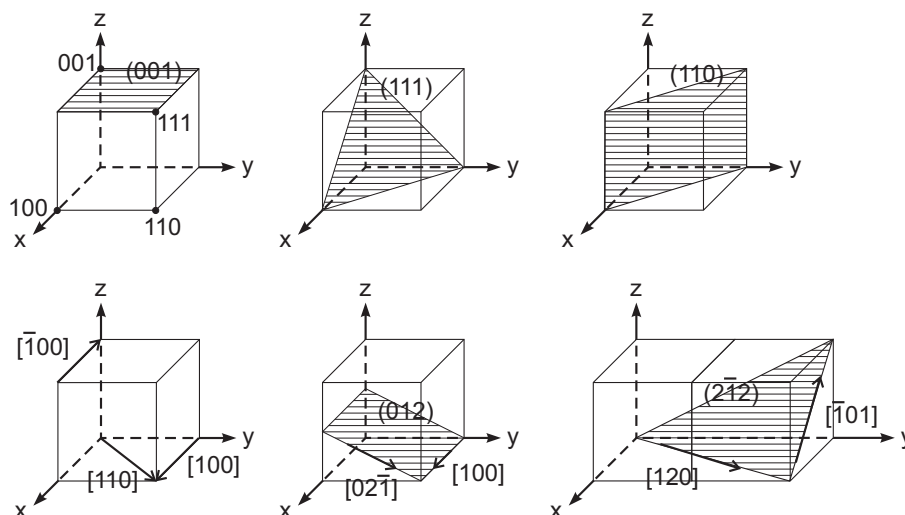
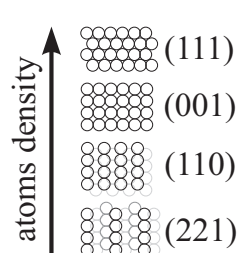


Figure 3.5: Lattice points coordinate, planes and directions in the cubic lattice of silicon.

A plane in the crystal is in turn identified by 3 indices (hkl) placed between parentheses which are obtained by considering the coordinate of the intersections between the plane and the 3 crystal axes. Actually the number h, k and l are always integers that are obtained by using the reciprocal of the intersection coordinates for each axes and reducing them to the smallest possible integers by clearing the common factors. If the integer is negative, it is represented by placing a bar on its top.

Three important crystal planes, the (100) plane, (110) plane and (111) plane have been illustrated in Figure 3.5. For example the (100) plane intercept the  $x$  axis in 1 (the reciprocal is  $1/1 = 1!$ ), and along  $y$  and  $z$  the 0 arises because in these cases the plane is parallel to the axis and thus will intercept it... at infinity - because we take the reciprocal of the intercept coordinate we get  $1/\infty = 0$ . Note that if the plane intercepts the axes at the origin, we need to use a parallel plane passing through neighbour cell to be able to compute the indices.



One of the cause of the anisotropy observed in crystal can be understood by considering the density of atoms 'at the surface' of a particular plane. Actually, let us use a stack of closely packed hard spheres as a simple 3D model for the atoms arrangement in a fcc lattice<sup>1</sup>. We observe different planes by cutting through this model<sup>2</sup>, and figured the atoms closest to the surface in black (and further away in lighter grey). We see that the (111) plane presents here the highest density of atoms

<sup>1</sup>This is not a Si crystal – here the unit cell has 1 atom only whereas in Si it has two.

<sup>2</sup>To perform this task the simplest is to use a computer system like the Surface Explorer proposed by K. Herman that was used to help produce this views <http://surfexp.fhi-berlin.mpg.de/>

possible in a plane with a closely packed hexagonal structure. The (100) plane present a square-packed arrangement, with more voids and thus a lower density of atoms, and other planes will be of different density.

---

**Example 3.1** Cleavage planes in  $\langle 100 \rangle$ -cut Si wafer.

---

**I**N  $\langle 100 \rangle$ -CUT WAFER of silicon, cleavage happens parallel to the main flat of the wafer which is located along the [110] direction which correspond for this cut to a (110) plane. Why doesn't it happen in the (100) or the (111) planes? In bulk crystal cleavage preferably happens parallel to high density plane. The density of atoms in a plane is found by counting the number of atoms belonging to the plane in one cell of the lattice and dividing by the surface of the plane in the cell.

For the (100) plane, by looking at the Si crystal structure we see that the atom at the center of the plane belong to the cell, while the four atoms at the corner belong to 4 other cells, the total number of atoms is thus:  $1 + 4/4 = 2$ , that is a density of  $\frac{2}{a \times a} = \frac{2}{a^2}$ .

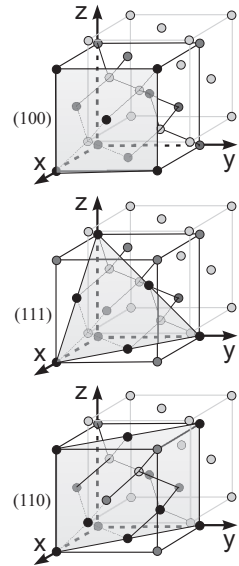
For the (111) plane, we see that the 3 atom at the corner of the triangle belong to 6 cells, while the 3 atoms at the edge belong to 2 cells, the total number of atoms is thus:  $3/6 + 3/2 = 2$ , that is a density of  $\frac{2}{\sqrt{2}a/2 \times \sqrt{3}a/\sqrt{2}} = \frac{4}{\sqrt{3}a^2}$ .

For the (110) plane we have the 4 atoms at the corners that belong each to 4 other cells, the 2 atoms in the top and bottom side diagonal that belong to 2 cells and the 2 atoms from the lighter fcc lattice that belong to the cell. The total number of atom is thus:  $4/4 + 2/2 + 2 = 4$ , that is a density of  $\frac{4}{\sqrt{2}a \times a} = \frac{2\sqrt{2}}{a^2}$ .

We have  $2\sqrt{2} > \frac{4}{\sqrt{3}} > 2$ , thus among these 3 planes the (110)-plane is indeed the preferred cleavage plane.

Note that the high density rules is not the only rule to decide for cleavage. Actually the cleavage plane is generally the plane with highest density... *that is perpendicular to the surface* for minimizing the cut cross-section and thus the energy required for the cut. In this way in a  $\langle 110 \rangle$ -cut wafer the cleavage plane generally will be the (111) plane, as the (110) plane is not normal to the surface.

---



The same indices are used to represent crystallographic direction as well. In this case the indices are obtained by considering the vector coordinate between two points of the lattice placed along the chosen direction. The coordinate are then reduced to the smallest set of integer and placed between brackets  $[hkl]$  to differentiate them from the plane indices  $(hkl)$ .

Interestingly, the direction normal to the  $(hkl)$  plane is the  $[hkl]$  direction. Moreover the angle  $\alpha$  existing between two directions is given by taking the dot product



between the two vectors:

$$\alpha = \cos^{-1} \frac{h_1 h_2 + k_1 k_2 + l_1 l_2}{\sqrt{h_1^2 + k_1^2 + l_1^2} \sqrt{h_2^2 + k_2^2 + l_2^2}}$$

In general, crystal symmetries result in different directions having the same physical properties and there is no need to distinguish them. In this case we use the  $\langle hkl \rangle$  notation to represent any of these equivalent directions, whereas for plane with equivalent orientation we use the  $\{hkl\}$  notation. For example, it is customary to give the crystallographic orientation of a silicon wafer by indicating the equivalent direction of the normal to the top surface. A  $\langle 100 \rangle$  wafer means that the normal direction is equivalent to the  $[100]$  direction, which could be  $[\bar{1}00]$  or even  $[0\bar{1}0]$ . In these different cases, the top surface of the wafer would be  $(100)$ ,  $(\bar{1}00)$  and  $(0\bar{1}0)$ , series of plane that present the same properties because of the silicon face-centered cubic lattice. For other lattices with less symmetries, care should be taken to use exact direction  $[hkl]$  to indicate the precise crystallographic direction of the top wafer surface.

### 3.2.2 Materials properties

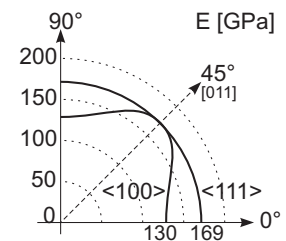
The choice of a good material for MEMS application depends on its properties, but not so much on carrier mobility as in microelectronics. Actually we select materials on more mechanical aspect: small or controllable internal stress, low processing temperature, compatibility with other materials, possibility to obtain thick layer, patterning possibilities... In addition, depending on the field of application, the material often needs to have extra properties. RF MEMS will want to be based on material with small loss tangent (for example high resistivity silicon), optical MEMS may need a transparent substrate, BioMEMS will need bio-compatibility, if not for the substrate, at least for a coating adhering well to the substrate, sensor application will need a material showing piezoresistance or piezoelectricity, etc. Actually, because the issue of material contamination is much less important in MEMS than in IC fabrication, the MEMS designer often tries to use the material presenting the best properties for his unique application. Still, from its microelectronics' root MEMS has retained the predominant use of silicon and its compounds, silicon (di)oxide ( $\text{SiO}_2$ ) and silicon nitride ( $\text{Si}_x\text{N}_y$ ). But actually, it was not purely coincidental because silicon, as K. Petersen explained in a famous paper [22], is an excellent mechanical material. Silicon is almost as strong but lighter than steel, has large critical stress and no elasticity limit at room temperature as it is a perfect crystal ensuring that it will recover from large strain. Unfortunately it is brittle and this may pose problem in handling wafer, but it is rarely a source of failure for MEMS components. For sensing application silicon has a large piezoresistive coefficient, and for optical MEMS it is transparent at the common telecommunication wavelengths.

Material	Young's modulus	Poisson ratio	Density
	GPa		
Stainless Steel	200	0.3	7900
Silicon (Si)	<100>130	0.25	2300
	<111>187	0.36	
PolySilicon (PolySi)	120-175	0.15-0.36	2300?
Silicon Dioxide (SiO <sub>2</sub> )	73	0.17	2500
Silicon Nitride (SiN)	340	0.29	3100
Glass	(BK7) 82	0.206	2500
	(SF11) 66	0.235	4700
Gold (Au)	78	0.42	19300
Aluminum (Al)	70	0.33	2700
SU8	4.1	0.22	1200
PDMS	0.0004-0.0009	0.5	970

Table 3.3: Material properties.

Of course, silicon being a crystal, it is anisotropic and its properties generally vary with direction. For example the elasticity modulus (Young's modulus) variation in the plane of the wafer are shown here for two different Si wafer cuts. We notice that due to crystal symmetries the <111>cut wafer presents an apparent isotropy for the in plane properties<sup>3</sup>, making it interesting for mechanical application. Other cuts will present a marked anisotropy, and for the <100>-oriented wafer we see that, depending on the direction in which a flexural beam is cut and bent, its elasticity modulus can vary between 169 GPa and 130 GPa.

Mathematically this behaviour can be taken into account by using a matricial formalism to manipulate tensors for relating physical quantities between one another. For example, elasticity theory can be used in an isotropic material to relate the different components of stress  $\sigma$  to the strain  $\epsilon$  using the stiffness matrix as  $\sigma = C\epsilon$  or the



<sup>3</sup>The value of 169 GPa in the graph may seem to contradict the Table where  $E=187.5$  GPa – but this last value correspond to the modulus in the <111>-direction, whereas in the graph we show the direction perpendicular to that direction ((111)-cut wafer)

compliance matrix as  $\epsilon = S\sigma$ :

$$\begin{bmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \\ \gamma_{XY} \\ \gamma_{YZ} \\ \gamma_{ZX} \end{bmatrix} = \begin{bmatrix} 1/E & -\nu/E & -\nu/E & 0 & 0 & 0 \\ -\nu/E & 1/E & -\nu/E & 0 & 0 & 0 \\ -\nu/E & -\nu/E & 1/E & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/G & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/G & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/G \end{bmatrix} \begin{bmatrix} \sigma_X \\ \sigma_Y \\ \sigma_Z \\ \tau_{XY} \\ \tau_{YZ} \\ \tau_{ZX} \end{bmatrix} \quad (3.1)$$

where  $\epsilon_I$  and  $\sigma_I$  represent the longitudinal strain and stress along direction  $I$ , and  $\gamma_{IJ}$  and  $\tau_{IJ}$  the shear strain and stress in the  $IJ$  plane<sup>4</sup>. In the case of unidimensional stress (e.g., along  $X$ ), the strain is simply described as  $\sigma_X = E\epsilon_X$ , where  $\sigma_X$  is the stress along the  $X$  direction,  $\epsilon_X = \delta x/x$  the strain along  $X$  and  $E$  the Young's modulus. We also observe the effect of the Poisson's ratio  $\nu$ , where the positive stress along the  $X$  direction does not only cause elongation along  $X$  as seen above but also contraction in the other direction as  $\epsilon_Y = -\frac{\nu}{E}\sigma_X$  (and similar along  $Z$ ).

For an anisotropic material, the relationship will not be as simple and the stiffness matrix (the inverse of the compliance matrix) will have a larger number of non-zero terms, coupling stress and strain components in a much more complex pattern:

$$\begin{bmatrix} \sigma_X \\ \sigma_Y \\ \sigma_Z \\ \tau_{XY} \\ \tau_{YZ} \\ \tau_{ZX} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ c_{12} & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ c_{13} & c_{23} & c_{33} & c_{34} & c_{35} & c_{36} \\ c_{14} & c_{24} & c_{34} & c_{44} & c_{45} & c_{46} \\ c_{15} & c_{25} & c_{35} & c_{45} & c_{55} & c_{56} \\ c_{16} & c_{26} & c_{36} & c_{46} & c_{56} & c_{66} \end{bmatrix} \begin{bmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \\ \gamma_{XY} \\ \gamma_{YZ} \\ \gamma_{ZX} \end{bmatrix} \quad (3.2)$$

The  $C$  matrix is symmetric (action-reaction principle) as the stress and strain tensors, limiting the number of independent terms. Actually instead of 81 terms<sup>5</sup> we have at most 21 independent terms as shown above.

Moreover, crystals have symmetries that may further limit the independent terms when we place ourselves in the crystallographic system of coordinate. In this

<sup>4</sup>We may note at this stage that we should have 9 components for stress or strain as they are second rank tensors, but for symmetry reason the shear  $IJ$  terms are equal to the  $JI$  terms, resulting in "only" 6 independent terms that are represented as a 6 terms vector. Using this trick, the 4<sup>th</sup> rank tensors  $C$  and  $S$  can be represented by a  $6 \times 6$  matrix.

<sup>5</sup> $C$  is a 4<sup>th</sup> rank tensor relating the two 2<sup>nd</sup> rank tensors of stress and strain, each with 9 components, thus requiring  $9 \times 9 = 81$  terms to relate them.

system of coordinate the  $X$ ,  $Y$ , and  $Z$  axes are, respectively, parallel to the  $A$ ,  $B$ , and  $C$  crystallographic axes that are defined by the atoms arrangement in the crystal. In cubic materials, such as silicon,  $A$  is the  $[100]$  direction,  $B$  the  $[010]$  and  $C$  the  $[001]$  and they form a Cartesian system of coordinate where the stiffness matrix has only 3 independent coefficient:

$$C = \begin{bmatrix} c_1 & c_2 & c_2 & 0 & 0 & 0 \\ c_2 & c_1 & c_2 & 0 & 0 & 0 \\ c_2 & c_2 & c_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & c_3 \end{bmatrix}$$

with  $c_1 = c_{11} = \dots = 166$  GPa,  $c_2 = c_{12} = \dots = 64$  GPa and  $c_3 = c_{44} = \dots = 80$  GPa for Silicon. If we compare this matrix to the isotropic case in Eq. (3.1), they seem similar, however, this form of the matrix in the case of Silicon is only valid for *one* particular  $X$ ,  $Y$  and  $Z$  Cartesian coordinate system parallel to the crystallographic ( $A,B,C$ ) axes – for isotropic materials it is true for *all* Cartesian coordinate system with any orientation. Alternatively instead of  $C$ , we could use the compliance matrix  $S$  and obtain:

$$S = \begin{bmatrix} s_1 & s_2 & s_2 & 0 & 0 & 0 \\ s_2 & s_1 & s_2 & 0 & 0 & 0 \\ s_2 & s_2 & s_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & s_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_3 \end{bmatrix}$$

with  $s_1 = s_{11} = \dots = (c_1 + c_2)/(c_1^2 + c_2c_1 - 2c_2^2) = 7.66 \cdot 10^{-12}$  Pa<sup>-1</sup>,  $s_2 = s_{12} = \dots = -c_2/(c_1^2 + c_2c_1 - 2c_2^2) = -2.13 \cdot 10^{-12}$  Pa<sup>-1</sup> and  $s_3 = s_{44} = \dots = 1/c_3 = 12.5 \cdot 10^{-12}$  Pa<sup>-1</sup> for Silicon. Instead of using the cumbersome matrix notation, it can be shown that for arbitrary direction in a cubic crystal we can define an equivalent Young's modulus in the direction  $(l_1, l_2, l_3)$  using:

$$\frac{1}{E} = s_1 - 2(s_1 - s_2 - 0.5s_3)(l_1^2l_2^2 + l_2^2l_3^2 + l_1^2l_3^2)$$

and a Poisson's ratio for any pair of orthogonal directions  $(l_1, l_2, l_3)$  and  $(m_1, m_2, m_3)$  using:

$$\nu = -E [s_2 + (s_1 - s_2 - 0.5s_3)(l_1^2m_1^2 + l_2^2m_2^2 + l_1^2m_3^2)]$$

---

**Example 3.2** Elastic modulus of <100>-cut Si wafer.

---

FOR THE CASE of silicon in a <100>-cut wafer (i.e. the <100>direction is perpendicular to the surface of the wafer), if we assume that the X-axis is normal to the wafer surface, the direction cosines in this (YZ) plane becomes simply  $(0, l_2, l_3)$ . Then the expression for the Young's modulus in the plane becomes:

$$\frac{1}{E} = s_1 - 2(s_1 - s_2 - 0.5s_3)l_2^2l_3^2$$

that is, by inserting the values of the material properties,

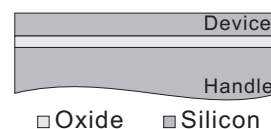
$$E = \frac{1}{s_1 - 2(s_1 - s_2 - 0.5s_3)l_2^2l_3^2} = \frac{1}{7.669 \cdot 10^{-12} - 7.107 \cdot 10^{-12}l_2^2l_3^2}$$

thus for the [010] and [001] direction we have either  $l_2$  or  $l_3$  that is zero and we obtain  $E = 130.3$  GPa. At  $45^\circ$  from these direction, we have  $l_2 = l_3 = \cos(\pi/4) = \sqrt{3}/2$  and  $E = 169.6$  GPa. These results corroborates the curves shown in the inset p. 74.

---

Additionally, silicon has a stable oxide easy to grow at elevated temperature, forming an amorphous film that is transparent and both thermally and electrically insulating. Actually this oxide has the smallest coefficient of thermal expansion of all known materials. Those properties are often put to good use during MEMS fabrication, where oxide support will be used to thermally insulate a pixel of a thermal camera for example.

A smart substrate based on silicon and coming from IC industry has made a remarked entry in the MEMS material list: the SOI (Silicon On Insulator) wafer. This substrate is composed of a thick silicon layer of several hundred  $\mu\text{m}$  (the handle), a thin layer of oxide of 1 or 2  $\mu\text{m}$  and on top another silicon layer, the device layer. The thickness of this last layer is what differentiates the IC and the MEMS SOI wafers: in the first case it will reach at most a few  $\mu\text{m}$  where in the later case, the thickness can reach 100  $\mu\text{m}$  or more. The high functionality of this material has allowed producing complete devices with very simple process, like the optical switch produced by Sercalo (discussed in more details in Section 3.6) fabricated with its mobile mirror, actuator and fiber alignment feature with one single process step!



Another interesting compound is silicon nitride ( $\text{Si}_x\text{N}_y$ ), which is stronger than silicon and can be deposited in thin layer with an excellent control of stress to produce 1  $\mu\text{m}$  thick membrane of several  $\text{cm}^2$ . A good control of stress is also obtained during deposition of poly-crystalline silicon. For example the Sandia National Lab's 'Summit V' process stacks five layer of poly-silicon allowing an unparalleled freedom of design for complex MEMS structure. Adding oxygen in

varying quantity to the nitride yields oxynitride compounds ( $\text{Si}_x\text{O}_y\text{N}_z$ ), giving the possibility to tune the refractive index between stoichiometric nitride ( $n=2.1$  @ 542 nm) and oxide ( $n=1.5$ ) - an interesting property for optical MEMS applications. Closing the list of silicon compound we can add a newcomer, silicon carbide SiC. SiC has unique thermal properties (albeit not yet on par with diamond) and has been used in high temperature sensor.

But silicon and its derivative are not the only choice for MEMS, many other materials are also used because they possess some unique properties. For example, other semiconductors like InP have also been micromachined mainly to take advantage of their photonics capabilities and serve as tunable laser source. Quartz crystal has strong piezoelectric effect that has been put into use to build resonant sensors like gyroscope or mass sensors. Biocompatibility will actually force the use of a limited list of already tested and approved material, or suggest the use of durable coating.

Glass is only second to silicon in its use in MEMS fabrication because it can easily form tight bond with silicon and also because it can be used to obtain biocompatible channels for BioMEMS. Moreover, the transparency of glass is what makes it often popular in optical MEMS application.

Polymers are also often used for BioMEMS fabrication where they can be tailored to provide biodegradability or bioabsorbability. The versatility of polymers makes them interesting for other MEMS application, and for example the reflow appearing at moderate temperature has been used to obtain arrays of spherical microlenses for optical MEMS. This thermoplastic property also allows molding, making polymer MEMS a cheap alternative to silicon based system, particularly for micro-fluidic application. Recently the availability of photosensitive polymers like SU8 [25] than can be spun to thickness exceeding 100  $\mu\text{m}$  and patterned with vertical sides has further increased the possibility to build polymer structure.

This quick introduction to MEMS materials needs to mention metals. If their conductivity is of course a must when they are used as electrical connection like in IC, metals can also be used to build structures. Actually, their ability to be grown in thin-films of good quality at a moderate temperature is what decided Texas Instruments to base the complete DLP micro-mirror device on a multi-layer aluminum process. In other applications, electroplated nickel will produce excellent micro-molds, whereas gold reflective properties are used in optical MEMS and nitinol (NiTi), presenting a strong shape memory effect, easily becomes a compact actuator.

Before we embark in the description of the fabrication processes used to shape and modify the raw materials, we think it is important to have a basic understanding of vacuum technology, a very important feature of most machines used for material deposition or structuring.

### 3.3 Vacuum technology

Many – if not all – of the machines used in microfabrication rely at one step or another on vacuum chamber. Actually, in the chambers used for wafer processing vacuum has many roles:

- emptying chamber from air or other contaminating gases, avoiding unintentional doping during material deposition or unwanted chemical reaction during etching
- controlling atoms density in the chamber, for igniting a plasma and increase atoms and molecules energy and reactivity
- sucking out reaction by-product and ensuring steady supply of gas onto wafer surface

In general these different functions are used conjointly or sequentially. For example during plasma enhanced chemical vapour deposition (PECVD - see Section 3.5.1.5), we first place the wafer in the chamber and removes air to prevent contamination. Then we introduce chosen ultra-pure gases and regulate the gas density for igniting a plasma. During the process we continuously pump out the by-product of the chemical reaction happening in the chamber. Finally, when the process is over, we empty again the chamber several time to remove any trace of harmful gases to protect the operator as it retrieves the wafer.

But, actually, what do we *really* mean with the word vacuum ? In general, this term refers to a space where the density of matter is vanishingly small. As a matter of fact, real vacuum (read complete emptiness) is hard to achieve in sizable volume, and even intergalactic space is not devoid of particles and there lies an estimated density of about 40 particles/m<sup>3</sup>. Of course, this is way smaller that the more than 10<sup>25</sup> particles/m<sup>3</sup> present at atmospheric pressure – but still not complete emptiness. Vacuum – hear, particle density – is unfortunately a quantity hard to measure directly and the level of vacuum is generally quantified by pressure, a measure of the force gas atoms or molecules exert on container walls. Actually gas atoms or molecules are moving at high speed as measured by their temperature<sup>6</sup> and if we look at their average kinetic energy we have<sup>7</sup>,

$$\frac{1}{2}m\bar{v}^2 = \frac{3}{2}kT$$

---

<sup>6</sup>Note that if they were not moving they would fall due to gravity and accumulate at the bottom of containers... and you would have to place your nose on the floor to breath.

<sup>7</sup>There is actually, according to the equipartition theorem, an energy of  $\frac{1}{2}kT$  per degree of freedom. For temperature only the 3 translations in space are considered, but note that polyatomic molecules will also be excited in rotation or internal vibration, but this does not change temperature nor pressure.

where  $v$  is the atom/molecule velocity,  $m$  their mass,  $T$  the absolute temperature and  $k = 1.381 \cdot 10^{-23}$  the Boltzman constant. This gives a RMS velocity of

$$v_{RMS} = \sqrt{v^2} = \sqrt{\frac{3kT}{m}}$$

that is about 500 m/s for common gases at room temperature. The high-speed particles hit the walls, and the change of direction (that is, the linear momentum) of the particle during this elastic collision comes from the impulse of the reaction force at the wall. Pressure is the sum of all this minute force from all the particles directed perpendicular to the surface and uniformly distributed over it ( $p = \Sigma F/A$ ). It will intuitively depend on the particles kinetic energy (mass and velocity) – which as we have shown above depends on the temperature – and their number per unit of surface (or density). Actually we have :

$$p = nkT \tag{3.3}$$

where  $n$  is the particle number density (in  $m^{-3}$ ). From this definition it is clear that at a given temperature, pressure is actually proportional to  $n$  and thus representative of vacuum – but care should be used when comparing ‘vacuo’ at different temperature, a better unit may then be to use  $n$  the number of particle per  $m^3$ . The pressure SI unit is a force per unit of surface or pascal, in honour of Blaise Pascal, a French physicist of the 17th century.

$$1N/m^2 = 1Pa$$

Unfortunately, pressure is probably the most abused units in the SI systems - nanotechnology is no exception, and although we will stick to Pa in this book, you’ll have to get used to bar, Torr or even – Queen forbids – PSI. The following list tries to make the best out of this historical mess that evolved from original need of hydraulic (working with m of water), to imperial and metric units, passing in between by mm of mercury or Torr:

- 1 Pa = 1 N/m<sup>2</sup> : the SI units
- 1 atm (standard pressure at sea level) = 101,325 Pa ( $\approx 10^5$  Pa) = 1013.25 hPa = 1013.25 mbar = 76 cm Hg = 760 Torr = 29.92 inHg = 10.33 m water
- 1 bar =  $10^5$  Pa = 750 Torr = 0.99 atm ( $\approx 1$  atm)
- 1 mbar = 1 hPa = 100 Pa = 750 mTorr
- 1 Torr (named after E. Torricelli) = 1 mm of Hg = 133 Pa
- 1 mTorr (often used in nanotechnology) = 1  $\mu$ m of Hg = 0.133 Pa
- 1 PSI (pound-force per square inch - used in aeronautics) = 6.89476 Pa (no kidding.)

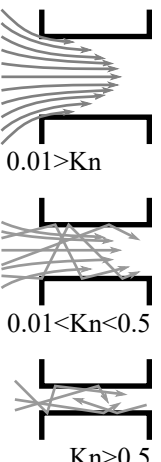


Formally, to get ‘vacuum’ the pressure needs to be below 300 hPa, that is the lowest pressure on the earth surface, at the mount Everest summit, where the thickness of air is the smallest and hence the hydrostatic pressure (i.e., the weight of the column of air above the surface) the lowest.

It is common to describe qualitatively vacuum from low to very-high, not by simply referring to a particular pressure value but by using a more useful quantification : the ratio between the dimension of the vessel where the gas resides and the distance that a gas molecule may travel without encountering another molecule. Actually, we use the Knudsen number  $Kn = \lambda/d$  that compares the width of the container (chamber or pipe)  $d$  with the gas molecules mean free path  $\lambda$ . The mean free path of the molecules is defined from gas kinetic theory as the mean distance that a molecule may travel between two consecutive collisions. It is inversely proportional to the density  $n$  and is given by :

$$\lambda = \frac{kT}{\sqrt{2}\pi p d_m^2} = \frac{1}{\sqrt{2}\pi n d_m^2}$$

where  $d_m$  is the molecular diameter. At atmospheric pressure ( $p \approx 10^5$  Pa), a nitrogen molecule ( $d_m = 0.37$  nm) will travel 62 nm between two collisions, while in high vacuum ( $p \approx 10^{-5}$  Pa) this distance will increase proportionally and become 510 m! This ratio is very important to describe vacuum as it governs the dominant collision mode when the gas molecules flow in a chamber : are the wall collisions or the inter-molecule collisions dominating?



Usually at atmospheric pressure, at one end of the regime ( $Kn = \lambda/d < 0.01$  : large channel), as the wall contact is infrequent and the inter-molecules collisions dominant we have viscous Poiseuille’s flow (cf. Sec. 4.3.3) (usually laminar because of low pressure and Reynold’s number). Then, as the interaction with the wall becomes dominant we pass through a transition regime (Knudsen flow) and have a progressive shift toward ( $\lambda/d > 0.5$  : narrow channel) molecular and diffusion flow. The equations governing these different types of flow show that for an elongated channel of diameter  $d$ , Poiseuille’s flow varies as  $d^4$  (cf. Eq. 4.10), molecular flow as  $d^3$  and diffusion flow as  $d^2$ , which could allow to experimentally find what regime is dominant in a particular case.

These different flow regimes are actually associated with a particular level of vacuum and we have :

- $Kn < 0.01$  : viscous flow and low vacuum (for  $p < 300$  hPa)
- $0.01 < Kn < 0.5$  : transitional (Knudsen) flow and medium vacuum
- $Kn > 0.5$  : molecular flow and high/ultra-high vacuum

Note that we often find in the literature a simple value of pressure for distinguishing between this different type of vacuum, but it is a simplification and the pressure alone is not enough to define the vacuum level, the dimension of the gas container is also of importance. For example, in a 1 cm wide pipe with a vacuum of  $p \approx 0.1$  Pa we are in high vacuum ( $Kn = \lambda/d = kT/\sqrt{2}\pi d_m^2 pd \approx 6.2$ ), the gas molecule may be considered ballistic as the mean free path is several cm and the molecule will rebound on the pipe walls more often that it will encounter another gas molecule. Still, we need to keep in mind that in a chamber of 1 m diameter, the pressure will need to be 100 times lower to reach the same level of vacuum and type of flow. Actually, for such chamber, we would have high vacuum for  $p < 0.01$  Pa, intermediate vacuum for  $0.01 \text{ Pa} < p < 2 \text{ Pa}$  and low vacuum above 2 Pa. In fact we obtain the same level of vacuum for the same  $pd$  product (provided we keep the same gas and the same temperature).

This last remark gives us a simpler way to estimate the level of vacuum: instead of using the Knudsen number we may use directly a scale in  $pd = kT/\sqrt{2}\pi d_m^2 Kn$  product, valid for a specific gas (an hypothetical 'air', very close to  $N_2$  and  $O_2$ ) at room temperature (25°C or about 300 K) :

- $pd > 0.6 \text{ Pa}\cdot\text{m}$  : low vacuum (viscous flow)
- $0.01 \text{ Pa}\cdot\text{m} < pd < 0.6 \text{ Pa}\cdot\text{m}$  : medium vacuum (transitional flow)
- $pd < 0.01 \text{ Pa}\cdot\text{m}$  : high/ultra-high vacuum (molecular flow)

A complete system to obtain vacuum is shown in Figure 3.6: basically we have a pumping system for sucking gases out of an hermetic chamber where vacuum will be established. We have also added other gas circuits for obtaining a simple process chamber: a viewport using a transparent window to observe the inside of the chamber, several regulated gas input for controlling gas introduction, a sensor for measuring the chamber pressure, which is generally linked to a variable valve placed before the pump for regulating outflow and chamber pressure.

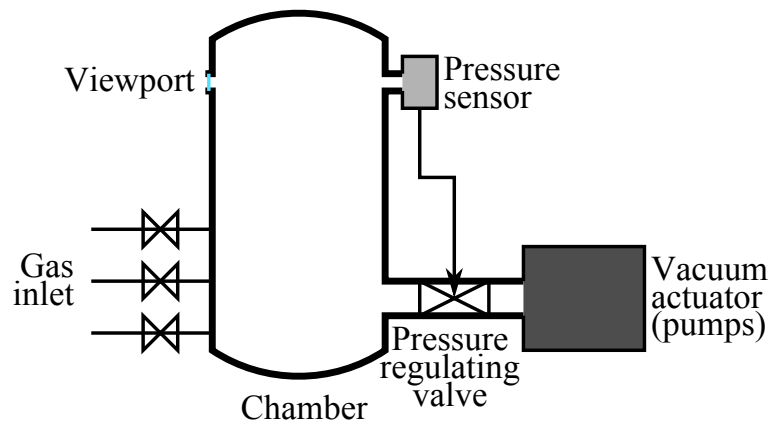


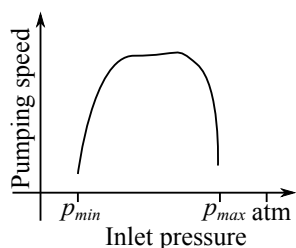
Figure 3.6: Typical vacuum elements in a process chamber.

The lowest vacuum that can be obtained in a chamber is fixed by the equilibrium arising between the gas removal rate by the vacuum actuator and the rate of gas entering the low pressure chamber. The main sources of gas entering the chamber are the leaks at the chamber port connections, the desorption of gas species at the inner surface of the chamber and even the permeation through the wall material by a diffusion like process. The problem of sealing at the detachable part (connecting hose, pump assembly, etc) is the most severe, and very-high vacuum chamber will mostly use welded or soldered assembly. However, for practical issues, most process chambers will use elastomer O-ring seals in groove for providing acceptable leak rate. They typically allow reaching vacuum below  $5 \cdot 10^{-6}$  Pa. For higher vacuum (down to  $10^{-7}$  Pa), the O-ring is replaced by a gasket of ductile metal like copper, pressed with large force between the two flanges of each side of the connection using clamping rings or for large diameter with multiple screws.

Ultimately, it is considered that a pressure of  $10^{-21}$  Pa resides in intergalactic space<sup>8</sup> (at a temperature of 3K, it means still about 40 particles per  $\text{m}^3$ ), but on earth, vacuum below  $10^{-10}$  Pa are hard to obtain and probably the best vacuum achieved so far is at the CERN between Switzerland and France where the Large Hadron Collider boast a vacuum of  $10^{-18}$  Pa - much lower than the “vacuum” present in the solar system !

### 3.3.1 Vacuum actuators

The generation of vacuum (that is pressure below earth atmospheric pressure) in a chamber is based on generic actuators called pumps. The pumps use different physical principle to evacuate gas from the chamber and they will work either by displacing the gas (using volume displacement or momentum transfer) or binding it to a fixed surface. The parameters of interest to describe the pump will be its maximum pumping speed (in  $\text{m}^3/\text{h}$ ), its maximum compression ratio (ratio between the pressure at inlet and at outlet ports) and sometimes the maximum differential pressure it can maintain between inlet and outlet ports.



We have listed some common pumps in Table 3.4, showing their most salient properties. The pumps are based on three principles : either a certain volume of gas is isolated and displaced from the inlet toward the outlet, or the gas molecule in the chamber are knocked away by collision with high kinetic energy elements, and we talk about momentum transfer, or finally the gas molecule are immobilized inside the chamber and bound to solid surface.

We note that many pumps can not work at atmospheric pressure ( $p \approx 10^5$  Pa) – generally the pumps used for high and ultra-high vacuum – meaning that a roughing pump will have to be used to lower the pressure in the chamber before the

<sup>8</sup>It may be 1 to 10 millions time bigger in galaxy and star system, but the meaning of pressure there is dubious as radiation pressure and solar wind may largely exceed the thermal force.

<b>Principle</b>	<b>Type</b>	$p$ <b>min</b> [Pa]	$p$ <b>max</b> [Pa]	<b>Backing</b>
Volume displacement	Rotary vane pump	$10^{-1}$	atm	no
	Roots pump (single stage)	$10^{-2}$	atm	yes
	Roots pump (multi-stage)	$10^{-2}$	atm	no
	Screw pump	$10^{-1}$	atm	no
Momentum transfer	Diffusion pump	$10^{-8}$	0.1	yes
	Turbomolecular pump	$10^{-10}$	100	yes
Gas binding	Cryo pump	$10^{-10}$	10	no

Table 3.4: Type of pump.

high vacuum pump may be used. Moreover, we see that some pumps requires a backing pump, that is, another pump at the outlet that evacuates the compressed gas and maintain the pressure there below the atmospheric pressure. With proper vanes and piping in the vacuum system, a single pump may be used alternatively as roughing and backing pump.

The pumping speed of a pump rarely gives the time it takes for emptying a chamber. The main cause is that the given pumping speed is a maximal value and it changes with the inlet pressure. Actually with decreasing pressure the pumping speed decreases and more generally it will follow a bell curve, as shown in the inset where we take the example of a high-vacuum pump that additionally can not reach atmospheric pressure. However, even taking this phenomena into account (and using a pump that remains efficient in the pressure range we want to reach) the time it takes to get the level of vacuum expected may be much longer than planned. The reason behind is because the rate assumes that the gas to be evacuated is *free to flow*. However, a lot of the gas to be removed is adsorbed on the surface of the chamber and pipes and takes a very long time to desorb, increasing the pumping time significantly – ultimate vacuum in chamber is obtained after hours or even days of pumping. Water vapour is particularly hard to desorb and some chamber walls are lined with heater to keep their temperature above 100°C, reducing the adsorption and helping the desorption of water. It is often said that the roughing pump is emptying the volume of the chamber whereas the high vacuum pump (usually using a backing pump) removes the gas from the surface. This is casually observed when, after a lengthy wait for obtaining a high vacuum, one introduces a little amount of (dry) gas in the chamber<sup>9</sup>: the pressure

<sup>9</sup>typically this is used for igniting a plasma that has difficulty to start because of the low pressure

risers rapidly, but decreases also very rapidly as soon as the gas valve is turned off, because the gas only goes into the ‘volume’ of the chamber but does not have time to adsorb on the walls. This also justifies the existence of chamber with a load-lock that is used to load the samples to the main chamber. In that case the main chamber is never brought to atmospheric pressure and little gas is adsorbed on its surface decreasing the time it takes to reach high vacuum before process can be performed.

### 3.3.1.1 Rotary vane pump

The rotary vane pump is based on volume displacement obtained by the rotation of an eccentric rotor with several adjustable vanes inside a circular chamber. The principle of operation is shown in Figure 3.7 with two vanes around the rotor : the gas is first sucked (phase 1 and 2) into a compartment defined by the front vane, which is then closed by a second vane (phase 3). The gas is then displaced toward the outlet port (phase 4) and is compressed (phase 5) opening the valve at the outlet port and being released at the higher pressure port. The pump is immersed

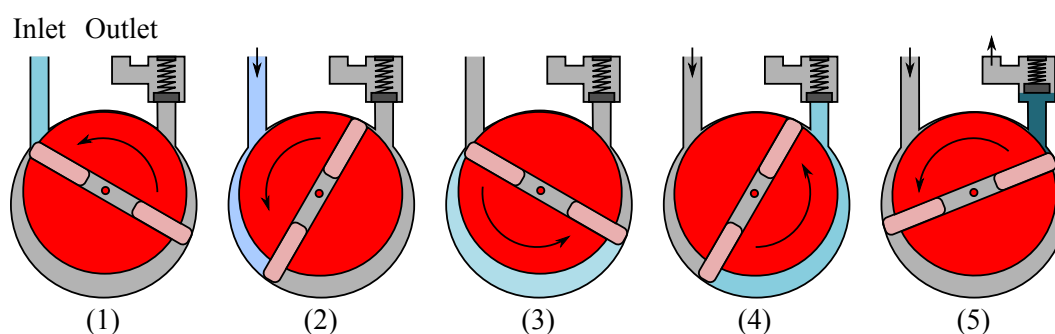


Figure 3.7: Sequence showing the principle of a operation of a rotary vane pump.

in oil that seals the contact region between rotor and stator, helps lubricate the whole pump and ensures heat removal. It is actually the main drawback of this type of pump, as the outlet port will see the oil which then migrates to the inlet port and toward other part of the vacuum system. Another problem may occur during the compression before the gas exits, as it may condense the vapour present in the gas, particularly water vapour. To avoid this it is possible to add a gas ballast that will reduce the pressure at the output port.

A single stage rotary vane pump may reach a pressure below 50 Pa and a two-stage pump will allow going down below 0.5 Pa.

### 3.3.1.2 Roots pump

The Roots pump takes its name from its inventors, the American brothers Philander and Marion Roots. The pump is based on volume displacement obtained with two counter-rotating rotors having interpenetrating lobes. The interest of

the design is the absence of friction (there is a gap between the lobes) and the balanced load around the rotating axis, allowing high rotating speed (5 krpm or more). The single stage Roots pump does not allow to obtain a large compression ratio ( $\approx 50$ ) and the pressure difference between its inlet and outlet port is at most 10 kPa, but they have a very high speed of pumping (from 200 m<sup>3</sup>/h to several 1000 m<sup>3</sup>/h). They are useful as roughing pump and with a backing pump they can reach pressure in the medium vacuum range.

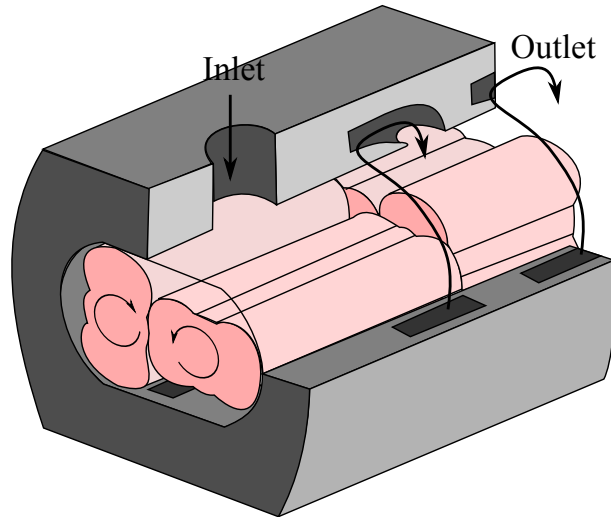


Figure 3.8: Cut-out view of the first two stages of a multi-stage Roots pump.

A multi-stage version of the Roots pump is actually shown in Figure 3.8, where the lower output from the first stage is funneled through a pipe in the housing to the the input of a second stage. In general there are 4 to 6 stages that can be placed in series. On the sketch, a shaft with two lobes is shown for simplicity, but in practice, 3 or 4 lobes are commonly used.

The multi-stage Roots pump may be used as a direct replacement of a rotary vane pump, with the advantage of the absence of oil in the gas path, lowering the chance of contamination. Another pump without oil is the screw pump that works with two intertwined Archimedes screws rotating in opposite direction in a principle similar to a multi-stage roots pump. Its compression ratio is lower than a multi-stage roots pump, but it may be used as a roughing or backing pump in many vacuum systems.

### 3.3.1.3 Scroll pump

The interest for dry pumps, that is pump without any lubricant in the gas path, has helped put in the forefront the scroll pump. Actually, although it was patented in 1905 by L. Creux a french inventor, the tight tolerance required for its operation did not allow building efficient models before the mid 1980's. as a good alternative for roughing. The working principle is based on volume displacement like roots

or rotary vane pumps. The scroll pump is built from two intertwined scrolls, one fixed and one mobile, having often both the profile of of an Archimedean spiral. As can be seen in Figure 3.9, by translating the mobile spiral along a circular path, the gas is trapped between the two spirals in a crescent section and progressively moved and compressed from the inlet port to the outlet port.

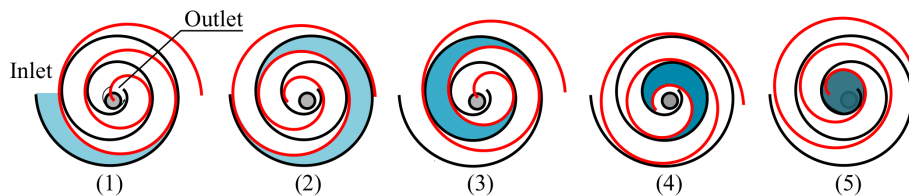


Figure 3.9: Principle of operation of the scroll pump.

The absence of reciprocating motion makes the pump silent and it is commonly used for fridge compressor. For process equipment the pump may pump from atmospheric pressure to about 10 Pa and often plays the role of roughing pump, or of backing pump for turbomolecular pumps.

#### 3.3.1.4 Diffusion pump

Historically many chambers were relying on diffusion pumps (Figure 3.10) for obtaining high vacuum. These pumps are based on momentum transfer obtained through the use of high speed vapour jet. At the bottom of the pump housing a heater is used to vaporize silicone oil and the vapour stream produced is diverted downward using jet assembly. This high-speed vapour flow pushes the residual gas molecules present in the pump chamber toward the bottom of the pump where they are pumped out by a backing pump. The oil vapour finally came into contact with the pump walls that are cooled (using water), condensing the oil that flows to the bottom of the pump to be evaporated again.

The main issue with this principle is the presence of... oil vapour. Actually vapour of oil can stream back toward the chamber or reach the outlet and baffles have to be placed at both ports to prevent this. Sometimes the inlet is equipped with a cold trap (cooled by liquid nitrogen) to condensate any oil vapour coming from the pump (and also some gas from the chamber). The simplicity and low cost of the diffusion pump made them popular, but they are relatively slow and often end up contaminating the chamber with trace of oil, preventing their use in high-vacuum system.

#### 3.3.1.5 Turbo pump

The turbomolecular pump or in short the turbo pump was patented by Pfeiffer Vacuum in 1958. Its operating principle relies also on momentum transfer, where rotating blades will knock gas molecules toward the outlet of the pump. Actually, as seen in Figure 3.11, the rotor blades are tilted and intertwined with stator

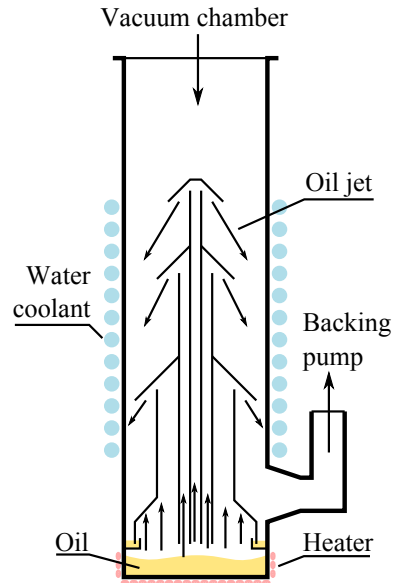


Figure 3.10: Principle of a diffusion pump.

blades tilted in a mirrored direction that channel the molecule toward the end of the pump. As usual for vacuum pumps, a better efficiency is obtained with multiple stage system and generally the rotor/stator pair of blades is repeated 5 to 10 times to increase the compression ratio than can be achieved.

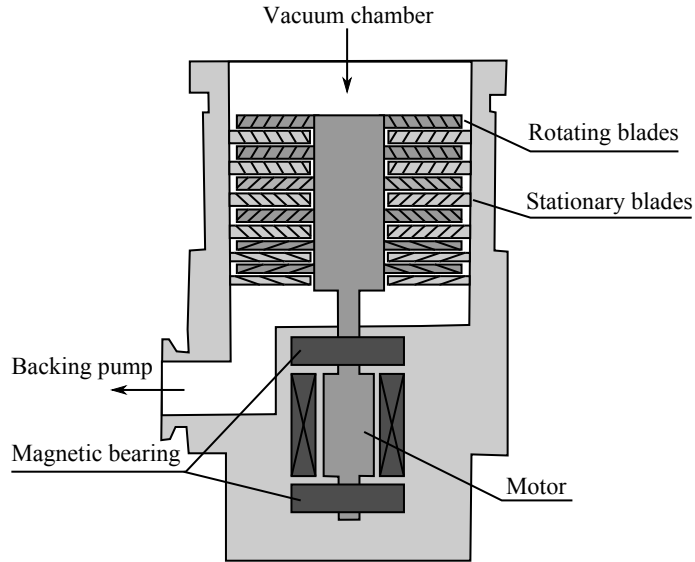


Figure 3.11: Principle of a turbomolecular pump.

The efficiency of the pump is related to its high speed of rotation and modern systems spins at several 10 krpm, even reaching 90 krpm. At these high speed, a key elements in the pump is the magnetic bearing that uses magnetic force to



levitate the rotor and to maintain it in place using complex feedback electronics. In this way there is no mechanical contact in the pump that requires no oil lubrication, providing an oil-free system suitable for high vacuum. The risk is that power failure would result in destruction of the pump as it brakes catastrophically while its bearing are no more controlled. However, in modern designs, the bearing are powered through the pump rotation itself, allowing the pump to stop safely.

Newer design will use combination of turbo pump followed by a drag pump. This pump consists in a rapidly spinning cylinder with a helical groove encased in a tight sleeve, that will again work by moment transfer and knock gas molecule toward the high pressure end. This hybrid pump can reach pressure down to  $10^{-9}$  Pa at high speed.

Still, as the momentum decreases with the molecule mass, so does the pump efficiency and it is usually difficult to pump hydrogen and helium efficiently with these type of pumps. For these gases, the compression ratio in a hybrid pump can be as low as  $10^4$ , compared to  $10^{10}$  for nitrogen.

### 3.3.1.6 Cryo pump

The cryogenic pump is based on trapping gas molecule using low temperature. Actually as the molecule touch a cold surface it get adsorbed, condensates and remains trapped there unable to escape.

It is clear that the capacity of such pumps is directly linked to the surface area of the cold trap, and accordingly large surface area are provided by using series of disk stacked together inside the pump housing. An example of the second stage of a cryopump with activated charcoal from an Edwards cryo pump is shown in Figure 3.12. In the cryopump the pumping happens in two stages: at the

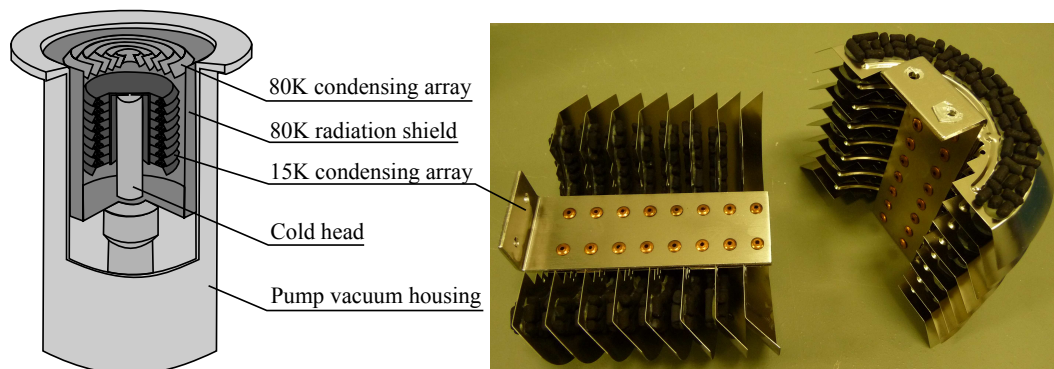


Figure 3.12: Cut-out schematic of a cryopump and picture of the porous charcoal array in the 15 K condensing stage.

entrance of the pump, series of disks maintained at a temperature around 80 K traps water molecules, while below, a second stage at a temperature  $< 15$  K is used for other molecules like nitrogen. The use of porous activated charcoal even allows for trapping hydrogen and helium in a process called cryosorption, although they

have an even lower condensation temperature. Actually hydrogen is not present in air and thus is not usually a problem, the ultimate vacuum reached by these pumps is actually directly linked to the atmospheric helium leaking rate, usually through permeation through the chamber materials.

For reaching the low temperature required inside the pump body, the system works similarly to a standard refrigerator, except that the compressed fluid inside the system is helium. The helium is compressed to a pressure of about 250 bar and flows into a motor driven piston assembly synchronized with the compressor. As the piston moves the helium gas expands and cools, then it returns and pushes the helium out of the cryopump and back to the compressor. The helium is re-compressed and the cycle repeats, removing each time a little heat until the cold head reaches about 10 K.

When all the sites for molecule adsorption have been taken on the cold surfaces, the pump needs to be regenerated. The pump is left heating and the gas molecules desorb from the surface and are sucked into the backing pump system.

Although they work at low pressure only, cryopump can be rather fast, with higher speed obtained by increasing the pump diameter. An extreme case is found in the ITER fusion experiment in France where the cryopumps used for maintaining vacuum in the tokamak are about 2 m diameter.

Because it has no moving parts the cryo pump boasts high reliability and systems have been sent to space for cooling low temperature detectors<sup>10</sup>.

### 3.3.2 Vacuum sensors

The measurement of pressure has been traditionally performed for centuries using the barometer proposed by Evangelista Torricelli – a collaborator of Leonardo da Vinci – at the beginning of the 16th century in Italy. Toricelli mercury barometer is measuring a difference of pressure between the inlet at pressure  $p_1$  and the end of the tube at pressure  $p_2$ . The difference in height  $h$  in the column of the liquid on the two branches of the U-tube is causing a difference of pressure  $\Delta p$ , the hydrostatic pressure. At equilibrium, this pressure has to be balanced by a difference of external pressure at both tube extremities. Considering that the liquid is incompressible, we can write the equilibrium of force on the column of liquid of section  $A$  as:

$$p_1 A + \rho h_1 A g = p_2 A + \rho h_2 A g \quad (3.4)$$

$$\Rightarrow \Delta p = p_1 - p_2 = \rho g h \quad (3.5)$$

where  $\rho$  is the liquid density,  $g$  the acceleration of gravity and  $h = h_2 - h_1$  is the difference in height of the liquid column on both side of the U-tube. The equation shows the interest of using mercury, its high density  $\rho$  making the column more compact – using water at atmospheric pressure would require a column more than

---

<sup>10</sup>Satellite can be rather hot as they generate heat internally, while cooling by heat radiation is very inefficient

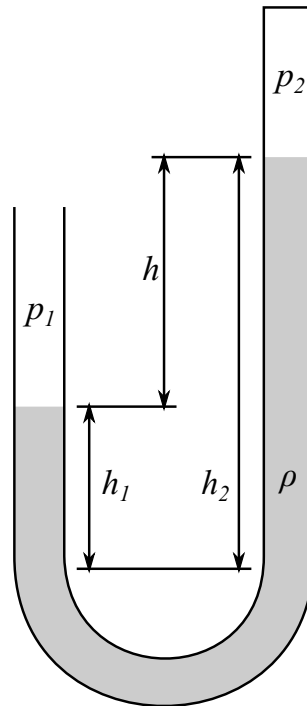


Figure 3.13: A simple barometer.

10 m high whereas we only need 760 mm of mercury. Notice that there is no complete vacuum at the back of the column, and  $p_2$  is larger than 0 because of the Hg-vapour and the presence of residual gas adsorbed on the glass surface. Actually early experimenters discovered that flashes of light appeared in this region when the barometer was moved: they were actually created by charges appearing by friction of Hg on glass and discharging in the low pressure gas present at the back of the column. The experimenters were actually witnessing, without knowing, the first artificial glow discharge (cf. Figure 3.20). When it is compensated for this effect, the mercury barometer remains one of the most precise measurement tool with a relative uncertainty of  $10^{-6}$ , making it still an important sensor for calibration standard.

For instrumentation, more practical sensors have been developed that may be split in two categories :

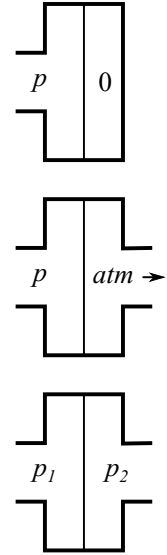
**direct measurement** the sensor (almost) directly measures the force applied by the gas over a surface, usually by recording the deformation of a flexible surface exposed to the pressure difference. The advantage of these techniques is that they are independent of the gas physical properties.

**indirect measurement** the sensor records one of the gas property that change with temperature (thermal conductivity, ionization field...) and from that result deduces the corresponding pressure. The advantage of these techniques is that they allow to measure very low pressure however they depend on the

gas being measured and their reading could be more ambiguous.

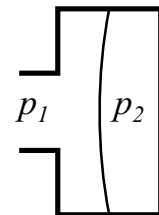
Direct pressure measurement is generally a differential measurement, recording a difference of pressure between two regions. We talk about absolute measurement when actually the reference is the 0 Pa pressure, barometric measurement when the reference is the atmospheric pressure and simply differential measurement when it records a difference between two regions in a system (often used for pumps).

When sensor are used to measure high pressure (compressed gas in bottle, compressed air...) the scale used is often the barometric scale and in that case to obtain the absolute pressure one should not forget to add the atmospheric pressure (about  $10^5$  Pa). The reason behind this choice is that for a vessel under high pressure, what really matters to estimate its resistance is the difference between the inside and outside pressures, which is actually the barometric pressure. Notice that it is intrinsically difficult for direct measurement methods to be truly absolute measurement method at very low pressure (we would need a chamber that remains at a calibrated pressure with an error inferior to the pressure that we want to measure) but indirect measurement have access to low pressure range more easily.



### 3.3.2.1 Membrane gauge

The membrane gauge is a direct measurement method where we measure the deformation of a membrane when there is a difference of force exerted on its two sides. One side is exposed to the pressure to be measured and the other side is at a reference pressure. The deformation of the membrane can be measured either by recording the stress induced inside the membrane or by measuring its deflection from the rest position. In the first case one can use resistors embedded in the membrane whose value will depend on the stress as described in Section 4.4.1 or use micromachined piezoelectric materials that would produce electric charges and ultimately current proportional to this stress. In the second case a popular technique will be to form a capacitor between the membrane and the backplate and to record the change of capacitance with the membrane deformation (Section 4.4.2) In both cases, the sensors are directly amenable to integration using MEMS technology and the different manufacturers are proposing sensors that can reach below  $10^{-2}$  Pa for temperature compensated capacitive gauge. One issue is that they work only over a bit more than 3 order of magnitude, requiring to pair sensors for operating from atmospheric pressure to the outset of high vacuum.



### 3.3.2.2 Pirani gauge

The Pirani gauge, named after its inventor M. Pirani at the beginning of the 20<sup>th</sup> century, is based on simple hot wire principle. When a filament is placed in a vacuum system and heated by Joule's effect, its equilibrium temperature depends on the number of molecule present in the system. Actually, as the pressure is decreased, there are fewer molecules coming in contact with the filament to carry away its energy, causing its temperature to rise. In the Pirani gauge, the filament used is made of metal (usually tungsten) whose resistance increases with temperature, and the pressure measurement becomes a simpler resistance measurement. This is usually accomplished using a Wheatstone's bridge configuration where one of the resistance is the W-filament exposed to the pressure to be measured. Sometimes, an identical compensating filament at constant pressure is placed in the other arm of the bridge.

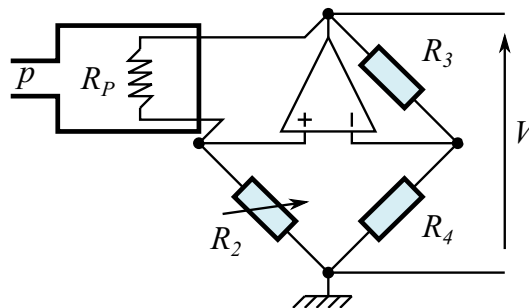


Figure 3.14: Control circuit for operating a Pirani gauge at constant temperature

The bridge may be operated at constant voltage or at constant current, but a particularly smart solution is to operate at constant gauge resistance by keeping the filament at constant temperature with the circuit shown in Figure 3.14. Actually in that case, the bridge is slightly unbalanced and the resistance of  $R_2$ ,  $R_3$ ,  $R_4$  are different from  $R_P$ . When the pressure decreases, the temperature of the filament increases causing  $R_P$  to increase. In turn this causes the bridge to unbalance and a voltage is developed across the bridge output terminals. The bridge control circuit senses the null voltage and decreases the voltage across the bridge until the null voltage is again zero. When the bridge voltage is decreased, the power dissipated in the sensor wire is decreased causing the resistance of  $R_1$  to decrease to its previous value. The opposite series of events happens for a pressure increase. The bridge voltage becomes a non-linear function of pressure that can be used using digital processing for providing an accurate reading of the pressure.

An interesting feature of this principle is that it can work over 6 to 7 order of magnitude providing sensors that can read from atmospheric pressure to the mid-high vacuum ( $10^{-2}$  Pa). Still in air, the oxygen present causes the filament to burn and the gauge can not be used at pressure above 1000 Pa in that case. For example, it should clearly be turned off - or isolated - before venting or opening a chamber.

### 3.3.2.3 Ionization gauge

This class of vacuum gauge uses an indirect measurement methods based on gas ionization. As an electron passes through a potential difference  $V$  it gains kinetic energy  $Ve$ , where  $e = 1.6 \cdot 10^{-19}\text{C}$  is the elementary charge. When this energy exceeds a critical value, corresponding to the ionization potential  $V_i$ , there is definite chance that collision between this electron and a molecule of gas drives out an electron and results in the formation of a positive ion. The number of gas ion produced by the accelerated electrons will be directly proportional to the density of the gas molecule <sup>11</sup>, and thus to the gas pressure (cf. Eq. 3.3). For monoatomic gases the ionization potential ranges between 3.88 V for cesium to 24.5 V for helium, while for diatomic gases ( $\text{O}_2$ ,  $\text{N}_2$ , ...) it is around 15 V. Ionization gauges are based on the measurement of the ion current, and differ only by the way the ionizing electrons are obtained. Higher electron energy allows obtaining more ions, however measurement shows that the maximum yield for most gases is obtained around an electron energy of 100 eV. Because ionization potential and efficiency varies with gas species, the reading of such gauge will always be influenced by the type of gas in the chamber – for example, ionization efficiency will be about 5 times lower for helium than for nitrogen. Accordingly, these sensors won't give accurate absolute pressure measurements unless they have been calibrated with the exact gas mixture that is to be measured.

The hot cathode gauge relies on thermionic emission of electron, where heating a conductive material gives the outer orbital electrons sufficient energy to overcome the work function barrier and to escape the filament forming a localized electron cloud. The electrons are then accelerated with a positively polarized electrode, hitting and ionizing residual gas molecules in the gauge. The resulting positive ions are collected by a third electrodes with a negative potential giving an indirect measurement of the vacuum.

The most successful design is the Bayard-Alpert gauge shown in Figure 3.15). For this gauge the filament will typically be biased to give thermionic electrons energetic enough to ionize any residual gas molecules with which they collide. The (positive) gas ion are then collected by the central electrodes held at a negative potential. For example, the grid (spiral electrode) is held at +150V while the central ion collector remains at -45V. the ion current is then proportional to the gas atoms density and thus pressure.

The Bayard-Alpert gauge improved on the original hot-cathode design where the ion and the electron electrodes positions were exchanged and the filament placed in the center. In this earlier configuration the ion electrodes collected an important flux of soft X-rays coming from the 150eV electrons hitting the grid and causing an important photo-electron current that is indistinguishable from the ion current. Even in the Bayard-Alpert configuration the gauge is limited to about  $10^{-11}\text{Pa}$  by the photo-electron current and if it is exposed to air at high pressure the filament

---

<sup>11</sup>Provided the density is so low that only one collision happens while the electrons goes from the anode to the cathode

simply burns.

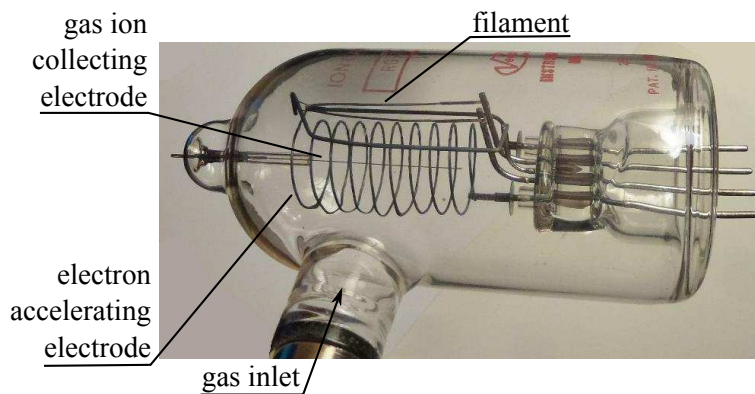


Figure 3.15: A typical Bayard-Alpert gauge.

For the cold cathode gauge, there is no source of thermionic electrons, and the plasma discharge starts from stray field emission (although the potential always remain under the field emission voltage a few events happen randomly), cosmic rays or radioactive decay. If the gauge is started at very low pressure, it may takes hours to obtain a stable discharge and usually the gauge is turned on at higher pressure. The electron is accelerated by the potential drop and deflected by a magnetic field that gives them an helical path toward the anode (+). In that way the distance traveled by the electrons is several hundred time longer than the real distance between the two electrodes, increasing the odds that the electron hits a molecule from the low pressure gas and ionizes it. Then the measurement of the ion current collected at the cathode (-) and amplified becomes an indicator of the pressure (hear density) of gas molecules.

A typical example of such gauge is the Penning gauge (or magnetron gauge),

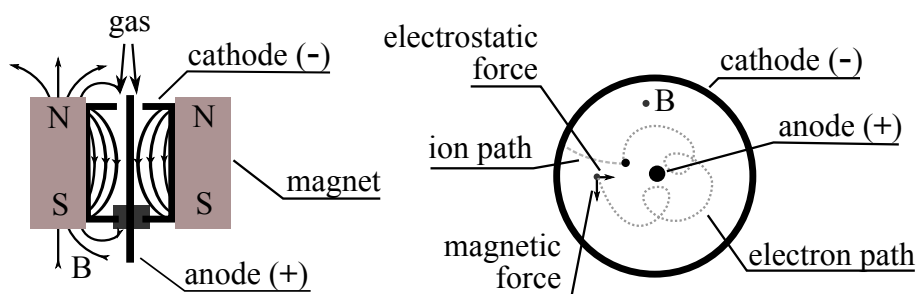


Figure 3.16: Side-view of a simplified inverted magnetron gauge (left) and top-view showing typical electrons and ionized gas molecules paths (right) .

but more modern design relies on the inverted magnetron design where the central electrode is held at a positive potential (anode) with respect to the ring electrode (cathode). In this configuration shown in Fig. 3.16 the discharge current is much more stable and it has now almost fully replaced the older magnetron designs where

the voltage polarity is inverted. Actually in real design the gauge uses a separate external polarizing cathode at the same potential as the ion collecting cathode, preventing false measurement due to field electrons and allowing to use voltage as high as 6 kV with a magnetic field of 0.2 T. Inside the gauge, the electrons will accelerate along a complex helical path dictated by the sum of the electrostatic ( $F_{ele} = q\vec{E}$ ) and magnetic ( $F_{mag} = q\vec{v} \wedge \vec{B}$ ) forces, whereas the ionized gas molecules, because of their much higher mass, will acquire a much lower velocity and thus won't see the magnetic field, before reaching the ring cathode.

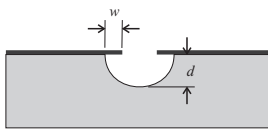
High quality inverted magnetron gauge may work with pressure down to  $10^{-13}$ Pa. It should be noted that the strong magnet used in the cold cathode ionization gauge will interfere with electrons and ions beams, and the position of the gauge should be carefully planned to avoid any adverse effect on the main machine process.

### 3.4 Bulk micromachining, wet and dry etching

Bulk micromachining refers to the formation of micro structures by removal of materials from bulk substrates. The bulk substrate in wafer form can be silicon, glass, quartz, crystalline Ge, SiC, GaAs, GaP or InP. The subtractive process commonly used to remove excess material are wet and dry etching, allowing varying degree of control on the profile of the final structure.

#### 3.4.1 Isotropic and anisotropic wet etching

Wet etching is obtained by immersing the material in a chemical bath that dissolves the surfaces not covered by a protective layer. The main advantages of this subtractive technique are that it can be quick, uniform, very selective and cheap. The etching rate and the resulting profile depend on the material, the chemical, the temperature of the bath, the presence of agitation, and the etch stop technique used if any. Wet etching is usually divided between isotropic and anisotropic etching (Figure 3.17). Isotropic etching happens when the chemical etches the bulk material at the same rate in all directions, while anisotropic etching happens when different etching rate exists along different directions.



However the etching rate never reaches 0, and it is actually impossible to obtain etching in only one direction. This is commonly quantified by estimating the underetch ( $w/d$ ), that is the lateral etching under the edge of the mask with respect to the vertical etching, as shown in the figure. This parameter may range between 1 for isotropic etching to about 0.01 for very anisotropic etch, obtained for example by etching Silicon in a KOH bath. For substrates made of homogeneous and amorphous material, like glass, wet etching must be isotropic, although faster surface etching is sometimes observed. However, for crystalline materials, e.g. silicon, the etching is either isotropic or anisotropic, depending on the



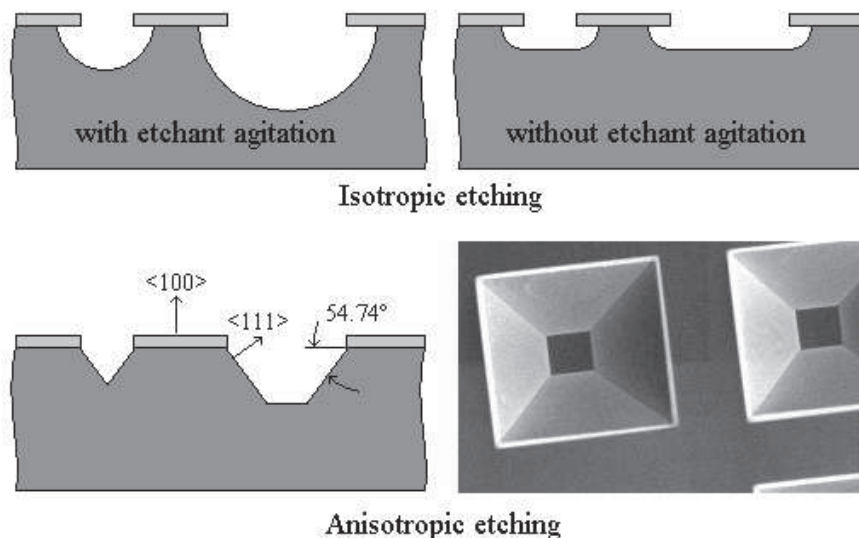
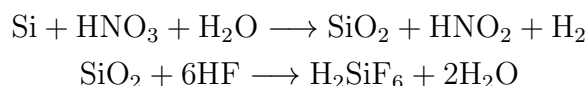


Figure 3.17: Isotropic and Anisotropic wet etching

type of chemical used. In general, isotropic etchants are acids, while anisotropic etchants are alkaline bases.

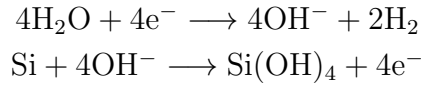
The top-left part of Figure 3.17 shows isotropic etching of silicon when the bath is agitated ensuring that fresh chemical constantly reaches the bottom of the trench and resulting in a truly isotropic etch. Isotropic wet etching is used for thin layer or when the rounded profile is interesting, to obtain channels for fluids for example. For silicon, the etchant can be HNA, which is a mixture of hydrofluoric acid (HF), nitric acid (HNO<sub>3</sub>), and acetic acid (CH<sub>3</sub>COOH). In HNA the nitric acid acts as an oxidant and HF dissolves the oxide by forming the water soluble H<sub>2</sub>SiF<sub>6</sub>. The two steps of the simplified reaction are:



The etching rate for silicon can reach 80 μm/min, and oxide can be used as mask material as its etch rate is only 30 to 80 nm/min. Etching under the mask edge or underetch is unavoidable with isotropic wet etching. Moreover, the etch rate and profile are sensitive to solution agitation and temperature, making it difficult to control the geometry of the deep etch usually needed for MEMS.

Anisotropic etching developed in the late 60s can overcome these problems. The lower part of Figure 3.17 shows features obtained by etching a (100) wafer with a KOH solution. The etched profile is clearly anisotropic, revealing planes without rounded shape and little underetch. Potassium hydroxide (KOH), tetramethyl ammonium hydroxide (TMAH) and ethylene diamine pyrocatechol (EDP) are common chemicals used for anisotropic etching of silicon.

For KOH and TMAH the simplified chemical reaction is written as :



We thus have a generation of hydrogen (bubbles escape during etching), and we notice that electrons are important elements in the reaction. Actually, the etching anisotropy has its roots in the different etch rates existing for different crystal planes that is generally thought to arise because of their different density of atoms and hence, of electrons. In fact, scavenging electrons will generally be a mean of stopping the reaction.

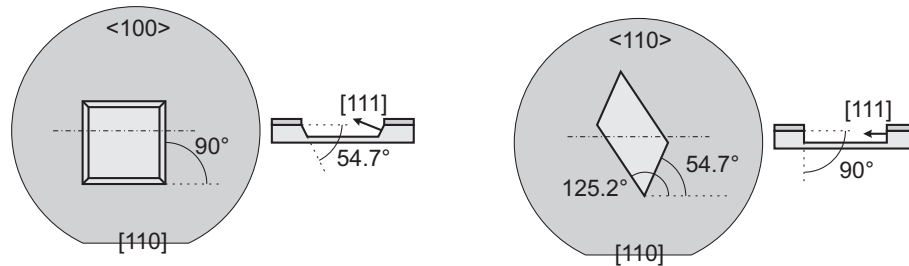


Figure 3.18: Orientation of the pattern edge for benefiting of the  $\langle 111 \rangle$  lateral etch stop plane in  $\langle 100 \rangle$  and  $\langle 110 \rangle$  wafers.

The anisotropy can be very large and for example, for silicon and KOH, the etch rate ratio can reach 400 between (100) and (111) planes and even 600 between (110) and (111) planes - meaning that when the etch rate for the (100) plane is about  $1 \mu\text{m}/\text{min}$  then the (111) plane will etch at only  $2.5 \text{ nm}/\text{min}$  effectively allowing to consider it as an etch-stop plane. With different combinations of wafer orientations and mask patterns, very sophisticated structures such as cavities, grooves, cantilevers, through holes and bridges can be fabricated. For example, if the (100) wafers in Figure 3.17 shows an angle of  $54.7^\circ$  between the (111) plane and the surface, typically producing V-grooves, (110) oriented wafer will present an angle of  $90^\circ$  between these planes resulting in U-grooves with vertical walls. To obtain these grooves, as shown in Figure 3.18, the mask pattern edges need to be aligned with the edge of the (111) planes. For a (100) wafer it is simple because the groove edge are along the  $\langle 110 \rangle$  direction, that is parallel to the main wafer flat. Moreover the four (111) planes intersect on the (100) surface at  $90^\circ$  and a rectangular pattern will immediately expose four sloping (111) planes and provide a simple way to obtain precisely defined pits or square membranes. (110) wafers are more difficult to handle, and to obtain a U-groove the side should be tilted by an angle of  $125.2^\circ$  with respect to the  $\langle 110 \rangle$  wafer flat. In addition, to obtain a four-sided pit, the two other sides should make a  $55^\circ$  angle with the flat direction - defining a non-rectangular pit that is seldom used for membranes.

If the control of the lateral etching by using the (111) planes is usually excellent, controlling the etching depth is more complicated. Monitoring the etching time

is the simplest technique. However this is limited by the etching uniformity in the bath, and by the variation of the etching rate. Actually, if the etching rate is known with 5% accuracy, after etching through a wafer of 300  $\mu\text{m}$ , the uncertainty on the etched depth is 15  $\mu\text{m}$ . We see that producing flat thin membranes of precise thickness (often  $t < 30 \mu\text{m}$ ) needed for pressure sensors will require a better approach than what can be achieved by this method. We have seen that we could use the self limiting effect appearing when two sloping (111) planes finally contact each other, providing the typical V-grooves of Figure 3.17. However, if this technique is interesting because it provides an etch stop and a structure of precise depth (we have  $d = w/\sqrt{2}$ ), it is unable to provide membrane with flat bottom. MEMS technologists have tackled this problem by developing other etch stop techniques that reduce by one or two order of magnitude the etch speed when the solution reach a particular depth.

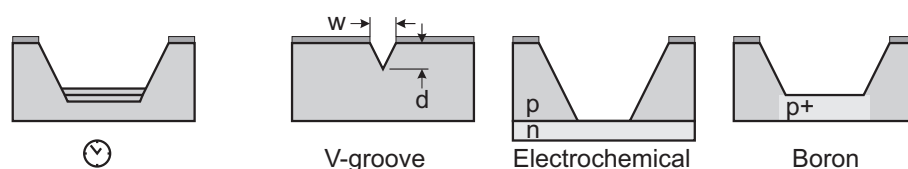


Figure 3.19: Comparison between timed etch and etch-stop techniques for controlling membrane thickness.

The electrochemical etch stop works by first creating a diode junction by using epitaxial growth or doping of a n-layer over a p-substrate. The junction is reverse polarized by contacting the substrate and the chemical bath, preventing current to flow between anode and cathode. As soon as the p-substrate is completely etched, a current can flow from the anode causing the apparition of a passivation layer by anodization, effectively stopping the chemical reaction and the etching. This process yields an excellent control over the final membrane thickness that is only determined by the thickness of the epitaxial layer, and thus can be better than 1% over a whole wafer.

Another popular method that does not require epitaxial growth, consists in heavily doping ( $> 10^{19} \text{cm}^{-3}$ ) the surface of silicon with boron by diffusion or implantation. As soon as the p+ doped zone is exposed the electron density is lowered, slowing down the etching reaction by at least one order of magnitude. However, if diffusion is used to obtain the boron layer, the resulting high boron concentration at the surface will decrease substantially the piezoresistive coefficient value making piezoresistors less sensitive. Ion implantation can overcome this problem by burying the doped layer a few  $\mu\text{m}$  under the surface, leaving a thin top layer untouched for the fabrication of the piezoresistors.

Actually, the seemingly simple membrane process often requires two tools specially designed for MEMS fabrication. Firstly, to properly align the aperture of the backside mask with the piezoresistor or other features on the front side (Figure 2.5) a double-side mask aligner is required. Different approaches have been

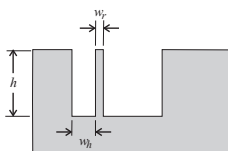
used (infrared camera, image storage, folded optical path...) by the various manufacturers (Suss Microtec, OAI, EVGroup...) to tackle this problem, resulting in a very satisfying registration accuracy that can reach  $1\ \mu\text{m}$  for the best systems. Secondly, etching the cavity below the membrane needs a special protection tool, that in the case of electrochemical etch stop is also used for ensuring the substrate polarization. Actually the presence of the cavity inevitably weakens the wafer and to avoid wafer breakage, the membrane is usually etched in the last step of the process. At that time, the front side will have already received metalization which generally cannot survive the prolonged etch and needs to be protected. This protection can be obtained by using a thick protective wax, but more often a cleaner process is preferred based on a mechanical chuck. The chuck is designed to allow quick loading and unloading operation, using O-ring to seal the front-side of the wafer and often includes spring loaded contacts to provide bias for electrochemical etch-stop.

The chemical used during anisotropic etching are usually strong alkaline bases and requires a hard masking material that can withstand the solution without decomposing or peeling. In general polymer (like photoresist) can not be used to protect the substrate, and if some metals (like tungsten) can be used effectively, in general a non-organic thin-film is used. For example, silicon oxide mask is commonly used with TMAH, while silicon nitride is generally used with KOH. Table 3.5 summarizes the characteristics of some anisotropic etching solution.

Of course anisotropic wet etching has its limitation. The most serious one lies with the need to align the sides of the pattern with the crystal axes to benefit from the (111) plane etch-stop, severely constraining the freedom of layout. A typical example is when we want to design a structure with convex corners - that is instead of designing a pit, we now want an island. The island convex corners will inevitably expose planes which are not the (111) planes and will be etched away slowly, finally resulting in the complete disappearance of the island. Although techniques have been developed to slow down the etch rate of the corner by adding protruding 'prongs', these structures take space on the wafer and they finally cannot give the same patterning freedom as dry etching techniques.

### 3.4.2 Dry etching

Dry etching is a series of subtractive methods where the solid substrate surface is removed by gaseous species.



Most of these methods are shared with microelectronics, but they take a different twist when they are applied to MEMS fabrication as in general MEMS necessitates deeper ( $> 2\ \mu\text{m}$ ) etching. As such we define an important parameter commonly used to describe dry etching: the aspect ratio. Actually we can define an aspect ratio for features ( $h/w_r$ ) and for holes ( $h/w_h$ ) with most technologies giving better results with features than with holes - but generally with only a small difference.

Solution	(100) Si etch rate ( $\mu\text{m}/\text{min}$ )	Etch rate ratio	Mask etch ( $\text{nm}/\text{min}$ )	rate	Boron etch stop ( $\text{cm}^{-3}$ )
KOH / H <sub>2</sub> O 44g / 100ml (30 wt.%) @ 85°C <sup>1</sup>	1.4	400 for (100)/(111) 600 for (110)/(111)	3.5 (SiO <sub>2</sub> ) <0.01 (Si <sub>3</sub> N <sub>4</sub> )		> 10 <sup>20</sup> rate/20
TMAH / H <sub>2</sub> O 28g / 100ml (22 wt.%) @ 90°C <sup>2</sup>	1	30 for (100)/(111) 50 for (110)/(111)	0.2 (SiO <sub>2</sub> ) <0.01 (Si <sub>3</sub> N <sub>4</sub> )		4 · 10 <sup>20</sup> rate/40
EDP (Ethy- lene diamine / pyrocatechol / H <sub>2</sub> O) 750ml / 120g / 240ml @ 115°C <sup>3</sup>	1.25	35 for (100)/(111)	0.5 (SiO <sub>2</sub> ) 0.1 (Si <sub>3</sub> N <sub>4</sub> ) ≈0 (Au, Cr, Ag, Cu, Ta)		7 · 10 <sup>19</sup> rate/50

<sup>1</sup> +largest etch rate ratio; -K ions degrade CMOS; -etch SiO<sub>2</sub> fast

<sup>2</sup> +SiO<sub>2</sub> mask; +CMOS compatible ; -large overtech

<sup>3</sup> +SiO<sub>2</sub> mask; +no metal etch; +CMOS compatible; -large overtech;  
-toxic

Table 3.5: Characteristics of some anisotropic etchants for silicon.

Typical values for this parameter would range between 1 (isotropic etch) and 50, for very anisotropic etching like the DRIE process.

As in many other processes, dry etching makes often use of a plasma. The plasma is an equal mixture of positive ions and high energy (=high speed) electrons with some neutral atoms that remains mostly electrically neutral. The plasma will help form a high quantity of reacting ions and radical, increasing the etching rate. In a plasma, new pairs of ion and electron are continuously formed by ionization and destroyed by recombination.



Plasma can be created by a glow discharge when a low pressure gas is submitted to a large electric field. In the glow discharge when the plasma is ignited, electrons close to the cathode are accelerated by the electric field across the Crooke's dark space until they reach the velocity necessary to ionize gas atoms. When the electrons hit a neutral gas atom they knock electrons out of their outer shell, ionizing them. The space region where this happen is called the negative glow and it is where most of the plasma is present. Actually the glow, whose colour (or wavelength) is characteristic of the plasma, comes from the photon emitted when a previously ionized atom and one electron recombine. The positively ionized atoms are in turn accelerated by the electric field in the Crooke's dark space, but because of their charge they move toward the cathode. When they hit it, the collision pro-

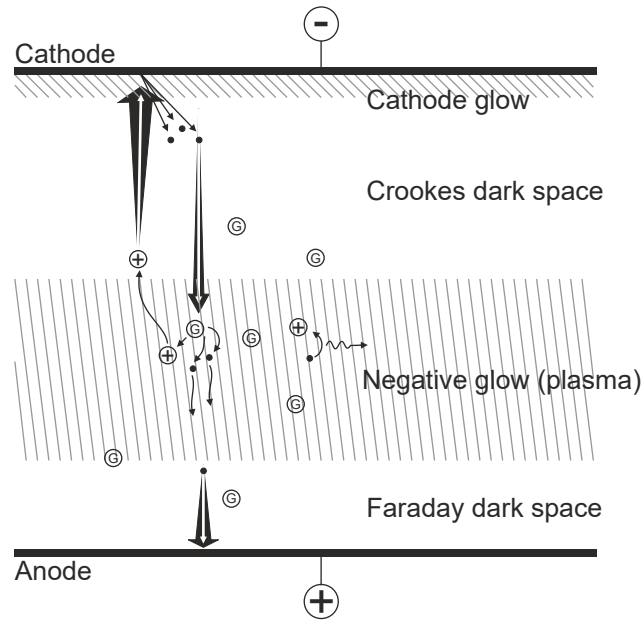


Figure 3.20: Glow discharge principle.

duces secondary electrons, which can be accelerated again to ionize more neutral atoms, creating a self-sustaining phenomena.

In the region after the negative glow, the electrons have lost most of their kinetic energy during the collision with the atoms and they accelerate through the Faraday dark zone before having enough energy to ionize atoms again and creating more glow zone in what is called the positive column. In general this part of the glow discharge region is not useful - except in neon tube, as it produces light emission - and the anode is kept close to the Faraday zone edge.

The pressure in the plasma has to be kept low so that the accelerated electrons do not encounter neutral atoms before they have acquired enough kinetic energy to ionize them. In the other hand if the pressure is too low the probability of the electrons ionizing an atom is very small and the plasma may not sustain itself. There is thus understandably a preferred range of operation pressure where the plasma is the brightest.

Due to the species present in the plasma, the etching can be obtained by three different mechanisms that achieve different results in terms of anisotropy or selectivity:

- physically by ion bombardment, knocking out atoms from the surface (ion etching or sputtering and ion-beam milling): anisotropic, non-selective
- by combining both physical and chemical effects (reactive ion etching or RIE): anisotropic, selective
- chemically by reaction giving gas by-products at the surface (plasma etching or radical etching): isotropic, selective

The dominant mode of operation will depend on the energy of the ions and the reactivity of the radicals. Usually the etching is more anisotropic (vertical) and less selective when it is more physical (corresponding to high energy plasma), while it is more isotropic and selective when it is more chemical (for low energy plasma). In the RIE mode, obtained for mildly energetic ions, the bombardment of the surface by ions allows to increase tremendously the rate of the chemical reaction, while keeping selectivity and providing anisotropy. For example, using an argon plasma (a non-reactive gas) in a chamber with  $\text{XeF}_2$  has been shown to increase the etch rate of Si by a factor of more than 10.

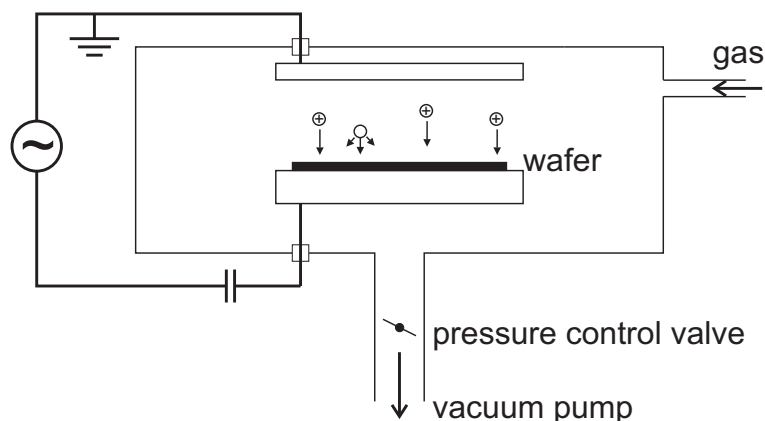


Figure 3.21: Capacitively coupled parallel plate RIE.

The RIE is the most versatile technique and is often used in MEMS. In its original configuration shown in Figure 3.21, if it is based on the glow discharge principle to generate the plasma, the excitation is obtained through a capacitively coupled RF source. Actually the high frequency RF source generates an alternating field at a frequency high enough (13.56 MHz) for affecting only the low inertia electrons. The electrons that are set into motion ionize the gas atoms, and end up on the chamber wall or on the electrodes. This loss of electrons in the plasma places it at a slightly positive potential. The electrons that come to the upper electrodes or the chamber wall are evacuated (ground), but those falling on the lower electrode accumulate, polarizing the plate with a negative voltage (a few -100 V). This self-polarization voltage will accelerate the positive ions from the plasma toward the wafer, resulting in RIE etching, an enhancement of chemical reaction due to the high energy of the ions. The directionality of the ions forced by the vertical electric field in the chamber enhances vertical etching over lateral etching, resulting in anisotropic etching profile. This simple RIE architecture has the main drawback that it is difficult to control separately the density of ions (ionization) and their energy (acceleration). Other scheme for RIE chamber have been designed (e.g. ICP-RIE in Section 3.6) that try to overcome this limitation, using additional electrodes and two RF sources to control independently these two parameters.

To improve the aspect ratio of the etching, several techniques based on RIE have been developed, usually trying to increase the anisotropy by protecting the sidewall during etching. For example, the SCREAM process developed in Cornell University alternate steps of etching and growth of oxide that remains longer on the sidewall, while for the RIE cryogenic process, very low temperature in a  $SF_6/O_2$  plasma is used to obtain continuous deposition of a fluoro-polymer on the sidewall during the etch. Finally the innovative Bosch process uses alternating cycle of etching and protection in the plasma chamber to achieve the same function. This important process is the cornerstone of DRIE micromaching and will be described in more details in a following section.

### 3.4.3 Wafer bonding

A review of MEMS fabrication technique cannot be complete without mentioning wafer bonding. Wafer bonding is an assembly technique where two or more precisely aligned wafers are bonded together. This method is often used simultaneously for device fabrication and also for its packaging - it belongs both to front-end and back-end process, another peculiarity of MEMS, but at this stage it is not surprising anymore!

Wafer bonding has the potential to simplify fabrication method because structures can be patterned on both wafers and after bonding they will be part of the same device, without the need for complex multi-layer fabrication process. The main issues that need to be considered to evaluate a wafer-bonding technique are : the bonding temperature (high temperature may damage the materials or structure on the processed wafer), the difference in coefficient of thermal expansion between the bonded materials (in the worst case causing debonding or affecting stress sensitive systems during use) and the permeability to gas and humidity of bond and bonded wafer (affecting long term reliability).

The bonding techniques are usually split between intermediate layer bonding technique, where an intermediate layer is used to form the bond between the two wafers, and direct bonding methods where there is no such layer.

Type	Bonding	Temp.	Stress	Hermeticity
Intermediate layer	epoxy	low	average	poor
	eutectic	average	average	very good
	glass frit	low	very good	very good
Direct	anodic	average	very good	excellent
	fusion	high	excellent	excellent

Table 3.6: Comparison of bonding techniques.



The simplest form of intermediate layer bonding is of course epoxy bonding. Although epoxy cures at low temperature ( $< 100^\circ$ ) and is cheap, such bonds pose performance problems as epoxy have large thermal expansion coefficient and are permeable to gas and moisture.

Intermediate-layer eutectic bonding is based on forming an eutectic alloy that will diffuse into the target wafer and form intermetallic to create a strong bond. For silicon-to-silicon bonding the intermediate layer is often gold which form a eutectic alloy with silicon at  $363^\circ\text{C}$ . Actually, if the two sides have been coated with gold, an even lower temperature can be used ( $\approx 250^\circ$ ) by applying pressure (50 MPa) at the same time. This process is called thermocompression bonding and was originally developed for gold wire-bonding in the late 1950s.

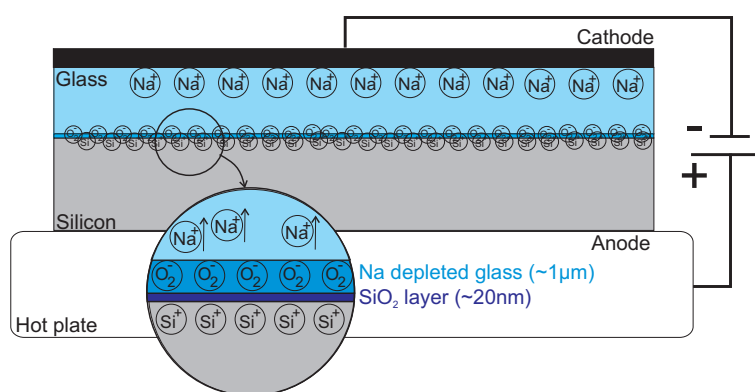


Figure 3.22: Anodic (field assisted) bonding principle.

The most commonly used MEMS bonding methods is probably the direct bonding method called anodic bonding (Fig. 3.22), or field assisted bonding, which is mainly used to bond silicon wafers with glass wafers. The technique works by applying a high voltage to the stacked wafers at elevated temperature ( $250^\circ\text{C}$ - $400^\circ\text{C}$ ) that induces migration of  $\text{Na}^+$  ions from the softened glass towards the cathode, leaving a negative space charge of oxygen ions. This ions repels free electrons in the semi-conductor, letting positive charge appear close to the interface. Accordingly, a strong electric field appears between silicon and glass, pulling both surface in intimate contact and, with the help of the elevated temperature, forming chemical bond with a thin  $\text{SiO}_2$  layer. The elevated temperature of bonding poses a unique challenge as it could introduce high stress when the bonded wafers are cooled down. Here, the choice of the material is of paramount importance : the Pyrex glass (Corning 7740) does not have the same coefficient of thermal expansion (CTE) than silicon<sup>12</sup>, but over the complete temperature excursion, the total strain in both material is kept about the same, insuring that the stress free condition existing at bonding temperature is also found at room temperature.

<sup>12</sup>Pyrex glass has a  $\text{CTE}=3.2 \cdot 10^{-6}/\text{K}$  constant until  $400^\circ\text{C}$ , while silicon has a CTE lower than Pyrex until about  $140^\circ\text{C}$  and higher above - the bonding temperature is chosen such that the integral of the difference of thermal expansion over the temperature range is close to 0

This technique is commonly used to fabricate sensors allowing for example to



Figure 3.23: Silicon pressure sensor SP15 bonded with glass cover (Courtesy Sensor AS - An Infineon Technologies Company).

obtain cavities with controlled pressure for pressure sensor as shown in Figure 3.23. At the same time, the glass wafer provides wafer level packaging, protecting sensitive parts before the back-end process.

The fusion bonding allows bonding directly two identical wafers (usually silicon-to-silicon), effectively achieving seamless bond possessing an exceptional strength and hermeticity. The surface are activated by plasma before the bond, but the technique still requires excellent flatness and high temperature, two hurdles that limit its use.

Before closing this section, it should be noted, that wafer bonding is also used to fabricate MEMS substrates such as SOI and SOG (silicon on glass) wafers. For SOI fabrication a thinned Si wafers is bonded to an oxidized Si substrate ( $t_{SiO_2} < 200$  nm) before it can be polished to the desired device thickness, resulting in the typical silicon-oxide-silicon stack.

### 3.5 Surface micromachining and thin-films

Unlike bulk micromachining in which microstructures are formed by etching into the bulk substrate, surface micromachining builds up structures by adding materials, layer by layer, on the surface of the substrate. The thin-film layers are typically  $1 \sim 5$   $\mu\text{m}$  thick, some acting as structural layer and others as sacrificial layer. Dry etching is usually used to define the shape of the structure layers, and a

final wet etching step releases them from the substrate by removing the supporting sacrificial layer.

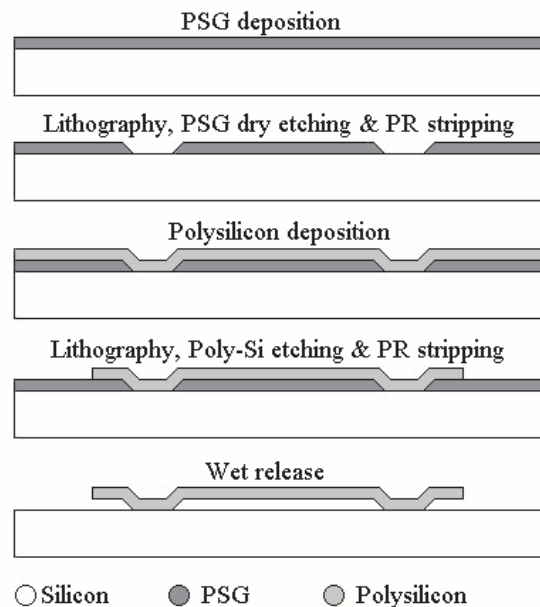


Figure 3.24: Basic process sequence of surface micromachining.

A typical surface micromachining process sequence to build a micro bridge is shown in Figure 3.24. Phosphosilicate glass (PSG) is first deposited by LPCVD to form the sacrificial layer. After the PSG layer has been patterned, a structural layer of low-stress polysilicon is added. Then the polysilicon thin-film is patterned with another mask in  $\text{CF}_4 + \text{O}_2$  plasma. Finally, the PSG sacrificial layer is etched away by an HF solution and the polysilicon bridge is released.

As a large variety of materials such as polysilicon, oxide, nitride, PSG, metals, diamond, SiC and GaAs can be deposited as thin film and many layers can be stacked, surface micromachining can build very complicated micro structures. For example Sandia National Laboratories is proposing a process with four polysilicon structural layers and four oxide sacrificial layers, which has been used for fabricating complex locking mechanism for defense application. Figure 3.25 demonstrates surface micromachined micro-mirrors fabricated using two polysilicon structural layers and an additional final gold layer to increase reflectivity. They have been assembled in 3D like a pop-up structure, using a micromanipulator on a probe-station.

### 3.5.1 Thin-film fabrication

The choice of the thin-film and its fabrication method is dictated by many different considerations: the temperature budget (limited by the maximum temperature

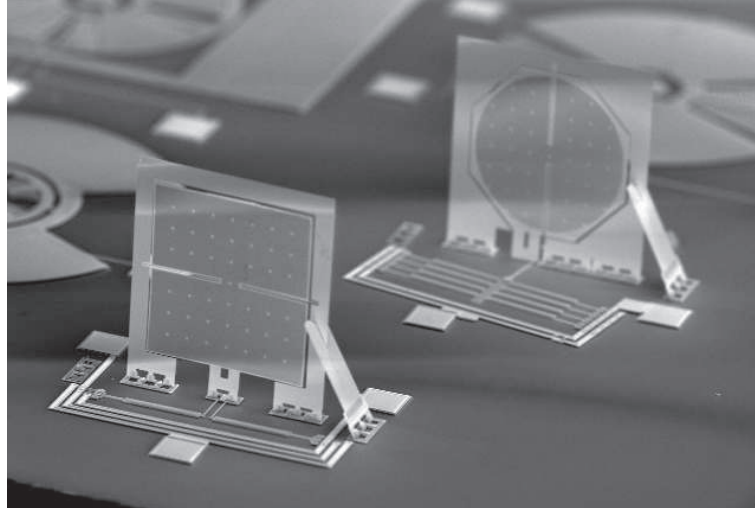


Figure 3.25: A micro optical stage built by surface micromachining.

that the substrate can withstand and the allowable thermal stress), the magnitude of the residual stress in the thin-film (too much stress cause layer cracking), the conformality of the thin-film (how the thin-film follows the profile of the substrate as shown in Fig. 3.26), the roughness of the thin-film, the existence of pinholes, the uniformity of the thin-film, the rate of fabrication (to obtain cost-effective thick thin-film)...

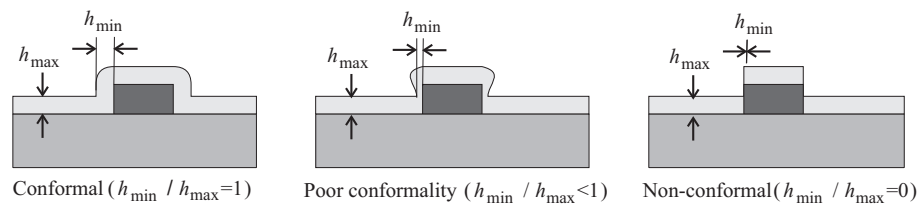


Figure 3.26: Conformality of layer deposited over a ridge.

Common thin-film fabrication techniques are the same as those used in microelectronics fabrication. Most of these methods are additive processes like chemical vapor deposition (CVD) at atmospheric (APCVD) or more often at low pressure (LPCVD), sputtering, e-beam or thermal evaporation, spin-coating, etc, but some methods are based on modifying process like oxidation (in dry or in wet oxygen), doping... These techniques have various characteristics that makes them desirable in different cases and are compared in Table 3.7. In general the best film quality is obtained at higher temperature, where the highest atom mobility assure good conformality and pin-hole free films with good adhesion. Moreover, the quality of a thin-film is linked with its microstructure that will heavily change if they are amorphous, polycrystalline or single crystal films, noting that single crystal thin-films can only be obtained with a few processes, like epitaxy.

Technique	Temperature	Conformality	Rate
Spin-coating	room temp.	--	++
Oxidation	very high	++	-
Evaporation	low	-	0
Sputtering	low	0	+
LPCVD	high	+	+

Table 3.7: Comparison of some thin-film fabrication techniques.

Besides the characteristics listed above, for surface micromachining we also need to consider an additional condition: the compatibility between sacrificial and structural layers. Actually, the selection of a suitable sacrificial material depends on the structural material and particularly on the availability of an etching method that can selectively etch the sacrificial material without significantly etching the structural materials or the substrate. A few common combinations of structural material and sacrificial etching method are shown in table 3.8, but the list is endless.

Structural material	Sacrificial material	Etchant
Polysilicon	Oxide(PSG, LTO, etc)	Buffered HF
Si <sub>3</sub> N <sub>4</sub>	Poly-Si	KOH
SiO <sub>2</sub>	Poly-Si	EDP/TMAH
Aluminum	Photoresist	Acetone/O <sub>2</sub> plasma
Polyimide	Cu	Ferric chloride
Ti	Au	Ammonium iodide
SiO <sub>2</sub> , Si <sub>3</sub> N <sub>4</sub> , metal	Poly-Si	XeF <sub>2</sub>

Table 3.8: Combination of materials and etchant for surface micromachining.

We will now give a detailed description of different thin-film processes.

### 3.5.1.1 Oxidation

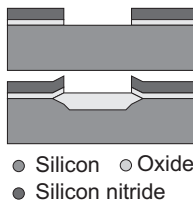
Oxidation belongs to the modifying processes, sharing with them a generally excellent conformality. Oxidation is a reactive growth technique, used mostly on silicon where silicon dioxide is obtained with a chemical reaction with a gaseous flow of dry or wet dioxygen. Using dry dioxygen results in a slower growth rate

than when water vapour are added, but it also results in higher quality films. The rate of growth is given by the well-known Deal and Groove's model as

$$d_o = \frac{A}{2} \sqrt{1 + \frac{t + \tau}{A^2/4B}} - \frac{A}{2},$$

where  $B$  is called the parabolic rate constant and  $B/A$  the linear rate constant that are obtained for the long and the short growth time limit respectively. Actually, for the short growth time limit, we notice that (neglecting the correcting factor  $\tau$ )  $\lim_{t \rightarrow 0} d_o = \frac{A}{2} (1 + \frac{1}{2} \frac{t}{A^2/4B}) - \frac{A}{2} = \frac{B}{A} t$ : we indeed have a linear growth rate with the slope  $B/A$ . Accordingly, we would find a parabolic approximation for long growth duration (cf. Problem 4).

Typical value for these constant at 1000°C are  $A = 0.165 \mu\text{m}$ ,  $B = 0.0117 \mu\text{m}^2/\text{h}$  and  $\tau = 0.37\text{h}$  in dry  $\text{O}_2$  and  $A = 0.226 \mu\text{m}$ ,  $B = 0.287 \mu\text{m}^2/\text{h}$  and  $\tau = 0$  in wet  $\text{O}_2$ . It should be noted that the model breaks down for thin dioxide ( $<300\text{\AA}$ ) in dry atmosphere because of an excessive initial growth rate that is modeled through the use of  $\tau$ . The passage from a linear growth rate to a parabolic rate is dictated by the need for the oxygen atoms to diffuse through the growing layer of silicon dioxide - the thicker the layer the slower its growth rate. The parabolic form of the growth rate is a typical signature of diffusion limited process. The diffusion rate is increased for wet oxidation as the presence of hydrogen facilitates oxygen diffusion through dioxide, resulting in much faster growth.



Actually, it is possible to control locally the oxidation by blocking in selected places the diffusion of oxygen to the surface, performing what is called a LOCOS process (local oxidation). In this process,  $\text{Si}_3\text{N}_4$  is deposited and patterned on the silicon surface (atop a thin oxide layer used for stress relieving) to act as a barrier against oxygen diffusion. During the oxidation

process in the furnace, oxide grows only in the bare regions while  $\text{Si}_3\text{N}_4$  prevents oxidation in the covered regions, effectively resulting in a patterned oxide layer. We note that the oxide growth happens in part below the original surface and that at the edge of the region, the oxide lateral growth lifts the nitride film resulting in a "bird's neck" profile of the oxide, both characteristics resolutely different from what would happen with etched oxide films.

For MEMS applications, an interesting feature of oxidation is that it results in a net volume change as the density is lower for oxide than for silicon. Actually, the volume increases by about 53% of the grown oxide. That is, for an infinite plane, the growth of a thickness  $d_o$  of dioxide results in the consumption of a thickness  $d_{\text{Si}} = 0.46d_o$  of silicon, but a net expansion of  $d_o - d_{\text{Si}} = 0.53d_o$  during growth. If this phenomenon may produce unwanted stress, it is also used to close holes in silicon or poly-silicon layers.

The relatively large range of variation of the oxidation growth rate, allows to obtain thick layer (up to  $2\mu\text{m}$  for 10h wet oxidation) or very thin layer (a few nm) with a good control. The thicker films can be used as mask for wet etching or as

sacrificial layer, and the thinner ones serve to produce nano-structures with high accuracy. This versatility and the high quality of the film produced give oxidation an important role in MEMS manufacturing.

### 3.5.1.2 Doping by diffusion and ion implantation

Doping is a process where impurities are introduced into a material to modify it, and as such belongs to the modifying processes. The impurities can be directly introduced during the fabrication of the substrate (for example, As can be introduced into the melt during the growth of the silicon ingot to obtain n-doped silicon wafer), but we are here interested by techniques that could be used selectively for doping locally and in-situ a thin layer of material. The two main techniques that are used are diffusion and ion implantation.

Diffusion is performed by placing the substrate in a high temperature furnace in presence of the doping species. In the main process currently used the doping species are present in gaseous form in the furnace or have been deposited as a thin film directly onto the substrate. The high temperature will agitate atoms strongly and allow the impurities atoms to move slowly inside the substrate, until temperature is lowered or their concentration is uniform.

Actually, the diffusion is governed by the Fick's laws which are derived from the statistical study of the random motion of particle due to thermal energy. The first law relates the flux of the diffusing species ( $\vec{j}$ ) with the gradient of concentration  $\overrightarrow{\text{grad}}C$ . The proportionality constant  $D$  is the diffusion constant, a material constant depending on the substrate, the diffusing atoms and the temperature.

$$\vec{j} = -D \overrightarrow{\text{grad}}C$$

which in system with one dimension gives  $\vec{j} = -D\partial C/\partial x$ . This equation translates the fact that the average flow of impurities will last until the concentration is equal everywhere and the gradient is null. By considering mass conservation equation in a small volume – that is no material is created locally but matters is only brought in by the diffusive flux – we may use the divergence operator to write that

$$\frac{\partial C}{\partial t} = \text{div}\vec{j}$$

Then, remembering that  $\overrightarrow{\text{div}}\overrightarrow{\text{grad}}\phi = \Delta\phi$  (the laplacian), we get Fick's second law that relates the evolution of the concentration with time:

$$\frac{\partial C}{\partial t} = D\Delta C$$

In one dimension problem it is given by  $\partial C/\partial t = D \partial^2 C/\partial x^2$ . These partial differential equations can be solved for different set of initial and boundary condition, depending on the diffusion configuration.

Of particular interest is the case of infinite source, where the concentration at the surface remains constant during the complete diffusion, as it happens during diffusion from a gaseous source in a furnace. In this case the concentration is given as:

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

where  $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = 1 - 2/\sqrt{\pi} \int_0^x e^{-\zeta^2} d\zeta$  is the complementary error function. From this equation we can obtain the diffusion depth  $d$  as:

$$d = 2\sqrt{Dt} \operatorname{erfc}^{-1} \left( \frac{C_d}{C_S} \right)$$

where  $C_d$  is the concentration in  $d$ . We verify here that the depth varies as the square root of time, that is, to go twice as deep, the diffusion needs to be 4 times longer. The surface concentration  $C_S$  remains constant during the diffusion and is generally given by the solubility limit of the impurities in the substrate (e.g.  $2 \cdot 10^{20} \text{at/cm}^3$  for Boron in Silicon at  $1100^\circ\text{C}$ ). Actually, the Fick's equations describe only the evolution of the concentration inside the material and not what happens at the interface between the substrate and the upper medium. The solubility limit gives the maximum concentration that is reached before the impurity forms clusters and small crystals which will require a lot more energy and generally does not happen. In this way it also gives the maximum impurities concentration that can be obtained by diffusion in a material.

Another common case is the case of finite source, where, after some time, the source is completely consumed by the diffusion, as happens when a thin-film that has been deposited on the substrate surface acts as the diffusion source. In this case, the relevant quantity is the total amount of dopant present  $Q$  that will finally fully diffuse in the substrate. In this case the solution to the Fick's second law is:

$$C(x, t) = C_S e^{-x^2/4Dt}$$

with in this case the surface concentration  $C_S = Q/\sqrt{\pi Dt}$ , decreasing steadily with the diffusion time. We obtain in this case a diffusion depth given by

$$d = 2\sqrt{Dt} \sqrt{\ln(C_d/C_S)}$$

varying again as a function of the square root of the time.

Ion implantation is comparatively a more versatile technique, but it requires a much more complex set-up (Figure 3.27). In this case the impurities are introduced into the substrate by bombarding the substrate with ions of the impurity traveling at high velocity. The ions will penetrate and be stopped after a short distance under the surface (at most a few  $\mu\text{m}$ ) by interaction with the electrons and atoms of the substrate material. The ion implanter is thus composed of a collimated source of ion, a section using high voltage to accelerate the ions, a mass spectrometer to sort the ions and only select the desired species and finally an electrostatic



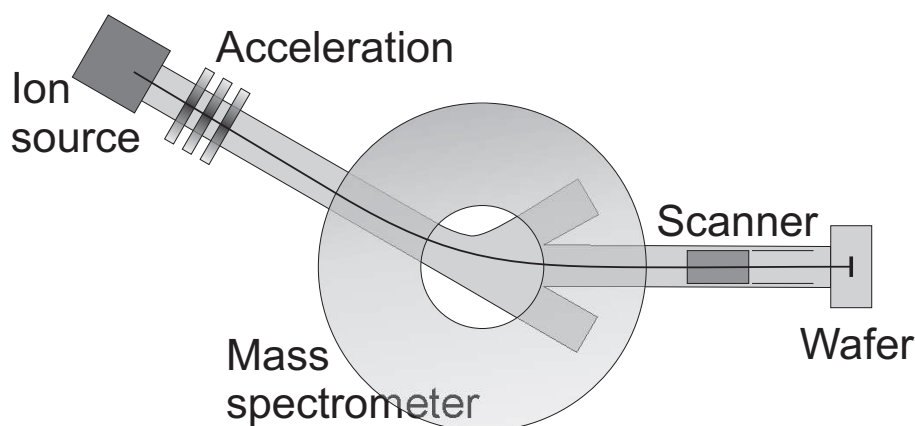


Figure 3.27: Schematic of an ion implanter.

scanning system allowing to direct the ion beam toward any place on the wafer surface.

The parameters governing ion implantation are simply the ion energy, expressed in eV and depending on the acceleration voltage, and the dose, that is the number of implanted atoms per unit area. The energy will determine the depth of implantation, with deeper implantation resulting in broader distribution of atoms. The dose is simply the ion current density ( $j$ ) multiplied by the implantation duration ( $t$ ) and divided by the charge of one ion. It can also be expressed by using the total current  $I$ ,

$$D = \frac{jt}{q} = \frac{It}{Aq}.$$

The implantation being a rather violent process, the collision of the impurity ions with the stationary atoms of the substrate cause them to recoil and generally results in amorphization of the doped portion of the substrate. In general recrystallization is needed and thus the implantation process needs to be followed by an annealing process at high temperature (800°C to 1200°C) under an inert atmosphere.

If ion implantation allows to tailor, by varying dose and energy, doping profile much more precisely than diffusion, it is not exempt of drawbacks. The implantation is a rather directional process and it is affected by shadowing behind tall structures and reflection on the side wall, making uniformity more problematic with high aspect ratio structures. Additionally, inside crystals there are often direction with less chance of nucleus collision were ions will be channeled much deeper. For example in Si, such phenomenon appears along the  $\langle 111 \rangle$  orientation, and when ion implantation replaced diffusion for microelectronics at the turn of the 1980's, the preferred Si wafer orientation changed from  $\langle 111 \rangle$  to  $\langle 100 \rangle$ . Finally, ion implantation is unable to provide deep doping profile ( $>$  a few  $\mu\text{m}$ ) and, contrary to microelectronics, diffusion remains widely used in MEMS fabrication.

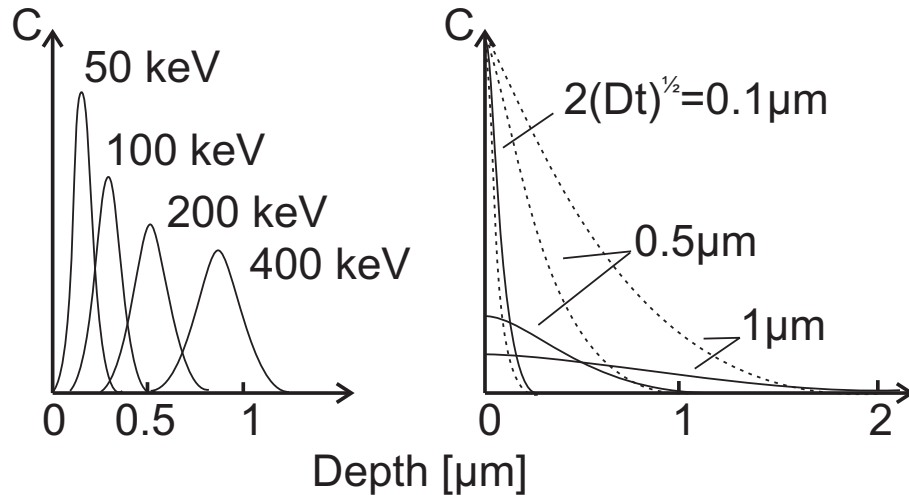


Figure 3.28: Concentration profile for (left) implanted ions of different energy (right) diffused atoms with different diffusion time and with (dots) infinite and (solid) finite source.

In general, it is possible to control locally the doping by placing a protecting layer over the zone that should not be doped before the doping process is performed. Actually the protecting layer will be doped at the same time as the exposed substrate, but will be removed in a later step leaving only the exposed substrate with doping. If the doping is obtained by diffusion, at the edge of the pattern lateral (isotropic) diffusion will occur enlarging the original pattern, while implanted layer will have more precisely defined edges.

### 3.5.1.3 Spin-coating

Spin-coating is a simple and fast technique that allows to deposit thin-films in liquid form. In general it is used to deposit polymer in a solvent, and particularly photoresist, but it can also be used for glass (spin-on glass) or other material (PZT...) with the sol-gel technique.

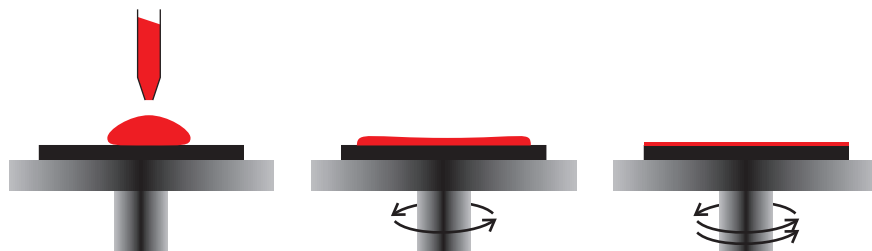


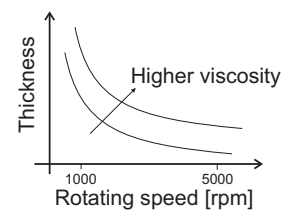
Figure 3.29: Spin-coating principle.

As shown in Figure 3.29, the substrate is mounted on a chuck that is able to spin at high speed before the liquid is dispensed in a puddle in the substrate

center. In a first spinning step at relatively low speed (a few 100 rpm), the liquid spreads over the entire substrate. Then a fast acceleration brings it to a high speed spinning for about 30 s where the layer reach its desired thickness. In general the high speed spin has to be kept between 1000 rpm and 5000 rpm to give optimum results, with increasing speed giving thinner films. The viscosity of the liquid is the main factor that determine the final layer thickness at a certain spinning speed, requiring a good temperature control (viscosity changes rapidly with temperature) to obtain reproducible results.

Photoresists exist in a wide range of viscosity, from a few cSt to 50000 cSt or more, allowing in a single spin to obtain thickness between about 100 nm and up to a few 100  $\mu\text{m}$ . A standard layer would be around 1  $\mu\text{m}$ , but thick resist are very attractive in MEMS fabrication. However, spin-coating very viscous polymer is difficult and will be helped by using a spinner with a co-rotating cover.

In that case the top cover above the substrate is rotating at the same speed, allowing to avoid the effect of air flow that creates uneven layer with particularly thicker edge called edge-bead. The edge-bead can be removed in a post-process conducted just after spin-coating, but their minimization is still important.



The method works best over a flat substrate as its conformality is not good. Actually, when the liquid is spun over a substrate with topography, the liquid surface tension will smooth out the sharp edge, rounding convex corner and filling concave ones. This phenomena can be put to good use to smooth out the topography of a wafer and has been use for local planarization. However in general this phenomena is a problem and a few methods have been developed to solve it in the case of photoresist where a good control of thickness is very important. One of such method is the spray coating method, where the photoresist is sprayed over the surface avoiding presence of thinner layer at corners. Photoresist has also been deposited by using electroplating. In that case a conductive substrate (i.e. metal coated) is placed in a special electrolytic bath and after application of a current, a layer of photoresist is grown uniformly on all the surface of the substrate, providing excellent conformality.

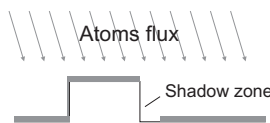


The main limitation with spin-coating is that the material need to be in liquid form, restricting the range of available material to a relatively small list. In general, as it is the case for photoresist, the material is dissolved in a solvent which is then evaporated after spin-coating leaving a hard film on the substrate. Alternatively, monomer can be spin-coated and then polymerized by heating or using UV exposure to form a film. Finally, in sol-gel method like the spin-on-glass (SOG), suspension can be spin-coated and will form a network upon thermal treatment.

### 3.5.1.4 Physical Vapor Deposition (PVD) techniques

There are two main techniques for physical vapor deposition : evaporation and sputtering. These techniques are low temperature and most often used to deposit conductive materials like metals (Au, Pt, Ti, Cr, Al, W, Ni...). However they can also be used for oxide or semi-conductive materials, resulting almost invariably in amorphous thin-films.

The principle of evaporation is very simple. The wafer is first placed with the source material in a vacuum chamber. The material is heated above its evaporation temperature sending vaporized atoms across the chamber. The atoms then condense on the colder surfaces : the wafer and the chamber walls. The material source is kept in a small crucible and heated using resistive heating (Joule's effect) or using an electron beam. In the later case, a high velocity beam of electron is directed towards the source. When the electron collide with the surface atom they transfer their kinetic energy, bringing the material to high temperature. The choice between the two techniques is not only governed by the evaporation temperature, which would make the e-beam technique superior as it can reach much higher temperature than resistive heating. Actually, some materials are easier to evaporate with one method or the other, and for example, some rare earth oxide can not be evaporated nicely with e-beam whereas they are successfully evaporated with resistive heating in a tungsten crucible.



For both heating techniques, the substrate has to be kept at a relatively large distance from the hot source to prevent the substrate from heating uncontrollably. However as the source material in the crucible has a relatively small size, the large distance result in a line of sight deposition with poor conformality. As we see in the figure, the horizontal surface will be uniformly coated, but the vertical surface will receive less material as their projected surface is smaller<sup>13</sup>. The vertical side can even be left in the 'shadow' and receive no material at all. This last problem can be generally avoided (except in narrow trenches) by rotating the substrate during the deposition. Of course the shadowing effect is not always a problem and actually lift-off process can benefit from it. Actually, if nothing deposits on the sacrificial material side this will facilitate its dissolution by providing a better access to the etching liquid.

The basic operation of the DC sputter relies on a plasma and is close to the glow discharge principle, but operated at a higher voltage. In that case, when the positive gas ions are accelerated in the Crooke's dark space and hit the target surface they have enough momentum to eject one or more atom of the target in addition to the secondary electrons. The neutral atom will then fly through the chamber and land onto the wafer.

The most commonly used gas for sputtering is argon, an inert gas, avoiding any

<sup>13</sup>This is the same phenomena that makes the temperature depends on the latitude on earth: at higher latitude the curve of the globe let the rays of the sun shine obliquely on the earth surface and thus make them illuminate a larger surface, bringing less heat there.

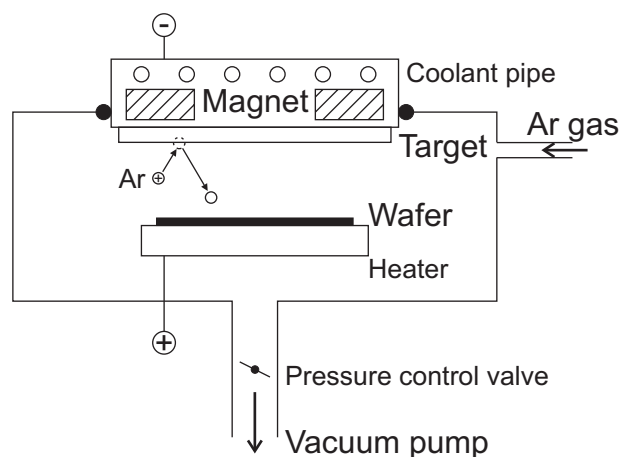


Figure 3.30: Typical DC Sputter schematic.

chance for the gas to chemically react with the substrate of the target. The target is located at the cathode which is biased at a few kV where the sputtering efficiency is highest for most materials. The substrate bias is also often set to a negative value ( $\approx -100$  V, that is, of course, much lower than the cathode voltage), to attract ions that will help compact the deposited film and remove the loose atoms. Alternatively a positive bias could be used to attract the electrons and repel the ions, resulting in different thin-films properties. Clearly, the substrate bias should not be made too negative otherwise sputter of the substrate will occur. In fact, this possibility is often used for a short time by biasing the substrate to 1 kV before the deposition is started in order to clean the substrate surface. Additionally, the wafer is normally heated at a temperature set halfway to the melting temperature ( $T \approx 0.5T_m$ ) to increase the atom mobility at the surface. This result in film having less porosity and lower stress. However, the main factor affecting stress is the gas pressure, and lower stress will be obtained with low pressure.

The magnet is used to increase the ionization yield of gas atoms. Actually the magnetic field traps the moving electrons (Lenz's law) in an helical path, increasing the length of their trajectory and thus increasing the chance they hit a gas atom. The yield increases dramatically ( $> 10\times$ ) resulting in a similar increase of the deposition rate. Most modern sputter are using this principle and are called magnetron sputter.

One issue with the DC sputter is that it can not be used efficiently for depositing insulating material. Actually, the charge building-up on the insulating target surface would finally repel the incoming ions and require a dramatic increase of the voltage to sputter the target material. However, instead of a DC potential, the sputter can also be operated with a pulsed field in the radio frequency (RF sputter) range (in general at 13.56 MHz). In this case the potential of the target is maintained negative for most of the period and then briefly turned positive. The short duration of the positive field does not modify much the momentum of the heavy ions that behave mostly as seen before in the DC field. In the other

hand the much lighter electrons are affected by the positive field and they move backward toward the target. This neutralizes the build-up of positive charge which would happen in DC sputter with non conductive target and allows deposition of insulating or semi-conducting materials.

The atoms coming from the target follows mostly a line of sight path, but the conformality is still better than with evaporation. Actually, as can be seen in Figure 3.30, owing to the large dimension of the target and its proximity with the wafer, target atoms arrive on the wafer within a relatively broad solid angle spread, decreasing the shadow effect usually observed with evaporation. Additionally, the atoms from the sputter have a higher velocity than atoms obtained by evaporation, resulting in layer with better adhesion to the substrate. Finally the deposition speed is higher with sputter, making it the tool of choice for metal deposition in the industry.

### 3.5.1.5 Chemical Vapor Deposition (CVD) techniques

Chemical vapor deposition techniques are split according to the operating pressure, from atmospheric pressure CVD (APCVD), low-pressure CVD (LPCVD) to finally ultra-high-vacuum CVD (UHCVD). However they all work on the same principle: the decomposition of a gas on the heated surface of the wafer in the absence of any reagent, a phenomena called pyrolysis. CVD is performed in a simple furnace with a gas inlet and connected to a vacuum pump to maintain a controlled pressure as shown in Figure 3.31.

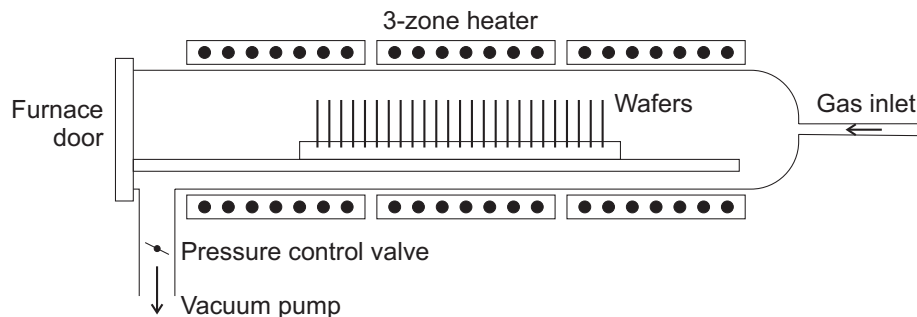


Figure 3.31: LPCVD furnace for thin-film deposition.

Depending on the gas, or gas mixture, used it is possible to deposit a wide variety of thin-films. The most commonly deposited films are polysilicon using the decomposition of silane ( $\text{SiH}_4$ ) at  $620^\circ\text{C}$ , silicon nitride using a mixture of dichlorosilane ( $\text{SiH}_2\text{Cl}_2$ ) and ammonia ( $\text{NH}_3$ ) at  $800^\circ\text{C}$  and low temperature oxide (LTO) using silane and oxygen ( $\text{O}_2$ ) at  $450^\circ\text{C}$ .

In general the resulting films are of good quality showing a very good conformality, having no pinholes with good deposition rate ( $1\mu\text{m}/\text{h}$ ). The stress in the film can be controlled by varying the temperature and the gas flow rate or composition. For example if stoichiometric silicon nitride ( $\text{Si}_3\text{N}_4$ ) will usually present a very

large tensile stress ( $>1$  GPa) making it crack easily for thickness above 200nm, increasing the amount of dichlorosilane in the gas mixture will result in silicon-rich silicon nitride films with a much lower stress that can become compressive and allow growing 1 $\mu$ m thick film and above. For polysilicon deposition using silane pyrolysis, increasing the temperature from 560°C to 620°C lowers the as-deposited stress, changing the compressive stress usually present in polysilicon films to tensile stress [23]. A subsequent high temperature ( $>950^\circ\text{C}$ ) anneal result in layer with very low stress, making Poly-Si the material of choice for building multi-layered structure on silicon surface.

Mixing of gas can result in other interesting variation in thin-films. For example it is possible to deposit oxynitride thin-films with an index of refraction anywhere between oxide ( $n=1.46$ ) and nitride ( $n=2.1$ ) by simply varying the ratio of oxygen and ammonia in a mixture with dichlorosilane.

The main concern with the LPCVD is the relatively elevated temperature needed for the pyrolysis of the gas. Actually, even the LTO deposition is actually rarely used as it shows a low conformality and usually oxide is deposited using pyrolysis of TEOS at 650°C. To circumvent this temperature problem, it is possible to use Plasma Enhanced CVD (PECVD). In this case a plasma is used to add kinetic energy to the thermal energy to facilitate the chemical reactions allowing the furnace to run at much lower temperature ( $<450^\circ\text{C}$ ). The low process temperature has made the PECVD reactor a popular tool for research and industry, used for many materials as shown in Table 3.9.

The PECVD furnace is completely different and more complex than its LPCVD counterpart, as in addition to temperature and gas control it needs the circuitry to excite and maintain a plasma inside the chamber above the wafer. As such, the deposition can only happen for one wafer at a time, and because of the horizontal position of the wafer, is more prone to contamination from falling particles. Additionally, the quality of the thin-films is usually lower - but adjusting the plasma parameter allows a better control on the mechanical properties of the film.

Actually LPCVD furnaces, even with vertically standing wafers, are also subjected to contamination from particles falling from the furnace walls. As we mentioned earlier the walls are hot and thus experience the same deposition as the wafer, accumulating across multiple run and finally falling on the wafer as particles. There are a few solution to this problem, and for example the LPCVD furnace may be oriented vertically instead of horizontally. Another possibility is to keep the wall cold so that no deposition occurs on them. This is not possible with LPCVD furnaces, but it can be achieved with Rapid Thermal Processing (RTP) furnaces. In a RTP furnace wafers are processed one by one horizontally, similar to what is done in a PECVD furnace. However here the high temperature is obtained using strong lamp whose NIR radiation is absorbed in the wafer. The walls can thus be kept at low temperature by water circulation preventing any deposition there.

It should be noted though, that for MEMS the problem of particles contami-

Material	Technique	Gas	T [°C]	Remark
Oxide (SiO <sub>2</sub> )	LPCVD LTO	SiH <sub>4</sub> + O <sub>2</sub>	425	low density
	LPCVD TEOS	Si(OC <sub>2</sub> H <sub>5</sub> ) <sub>4</sub>	700	good film
	PECVD	SiH <sub>4</sub> + N <sub>2</sub> O	300	contain H
Nitride (SiN <sub>x</sub> , Si <sub>3</sub> N <sub>4</sub> ...)	LPCVD Si <sub>3</sub> N <sub>4</sub>	SiH <sub>2</sub> Cl <sub>2</sub> + NH <sub>3</sub>	800	high stress
	LPCVD SiN <sub>x</sub>	SiH <sub>2</sub> Cl <sub>2</sub> + NH <sub>3</sub>	800	low stress
	PECVD SiN <sub>x</sub>	SiH <sub>4</sub> + NH <sub>3</sub>	300	contain H
Silicon (PolySi, a-Si...)	LPCVD PolySi	SiH <sub>4</sub>	620	small grain
	LPCVD a-Si	SiH <sub>4</sub>	570	amorphous
	PECVD a-Si	SiH <sub>4</sub>	280	contain H
Tungsten (W)	LPCVD W	WF <sub>6</sub> + SiH <sub>4</sub>	440	good conf.

Table 3.9: CVD of common thin-films.

nation is not as bleak as it could seem because the dimension of the feature are usually much larger than for ICs. Actually fabrication process generally generates much more small particles than big ones: in a typical cleanroom air, when the size of particles goes from 1 $\mu$ m to 0.01 $\mu$ m, the particle density may increase by 2 order of magnitude! As such, if we accept the rule of thumb that particle should be smaller than about one third of the critical dimension in the layout, we understand easily that MEMS with feature size rarely below a few  $\mu$ m are much less impacted by the particle contamination than ICs with 50nm design rules. We can also remark that the expected gain in yield achieved by shrinking the chip dimension (smaller area has less chance to see one particle) can be easily offset by the dramatic increase of smaller particle number noted above: to keep the same yield, the environment should actually be cleaner, and the number of smaller particle should decrease proportionally with the scaling factor.

### 3.5.1.6 Epitaxy

Epitaxy is a CVD techniques, as it generally relies on furnace very similar to what is used for LPCVD, but actually it presents features that makes it unique. The main difference between epitaxy and other CVD techniques is that in the case of epitaxy the structure of the thin-film depends on the substrate, and particularly, epitaxial growth allows to obtain single crystal layers. Actually if in the case of CVD the deposition is relatively random and independent of the substrate, generally resulting in amorphous or polycrystalline films, with epitaxy the thin-film will grow in an ordered manner determined by the lattice of the substrate.



If the material of the epitaxial layer is the same as the substrate, the process is called homoepitaxy and heteroepitaxy otherwise. Moreover, depending on the match between the lattice period of the substrate and the film, we can distinguish three types of epitaxial growth:

**commensurate growth** when the substrate and the layer have the same crystal structure and lattice constant,

**incommensurate growth** when they don't have the same lattice constant resulting in point defects at the interface,

**pseudomorphic growth** when they don't have the same lattice constant but the epitaxial layer strains to match the lattice of the substrate.

The growth of high quality crystals, like silicon, is generally obtained by the Czochralsky method, which consists in pulling slowly from a melt very large single crystal starting from a small seed crystal. Actually, this method could be described as an extreme case of homoepitaxy with commensurate growth.

The epitaxy process needs a furnace similar to a LPCVD furnace, but in practice the process is relatively more complicated to control. On silicon the main process is based on the reduction in a  $H_2$  atmosphere of  $SiCl_4$  at  $1200^\circ C$  with HCl as a by-product. However the high temperature makes it hardly useful, except as a first process step, and lower temperature process using dichlorosilane above  $950^\circ C$  have been developed, but are harder to control resulting often in polycrystalline layers.

The main interest of the technique is in the high quality of the grown layer, which results in good electronics properties, important for optoelectronics (solar cell, laser diode...) and some specific electronics circuits, and good mechanical properties (low stress) more interesting for the MEMS application. The relative difficulty of the technique makes it rarely used in MEMS fabrication, with the notable exception of the process used by Bosch for their multi-user foundry process. In the MPW process the structural layer is a  $10.5\mu m$  polycrystalline layer grown by epitaxy (called epipoly). In this case the interest is the growth speed (that could exceed  $0.3\mu m/min$ ) that can be obtained without sacrificing the low stress present in the layer.

### 3.5.2 Design limitation

The flexibility of surface micromachining is not free of unique problems that need addressing to obtain working devices.

During layer deposition, a strict control of the stress in the structural layer has to be exerted. Compressive stress in a constrained member will cause it to buckle, while a gradient of stress across a cantilevered structure causes it to warp, resulting in both case in probable device failure.

The possibility to stack several layers brings freedom but also adds complexity.

Actually there is large chance that the topography created by the pattern on underlying layer will create havoc with the upper layer, as illustrated in Figure 3.32.

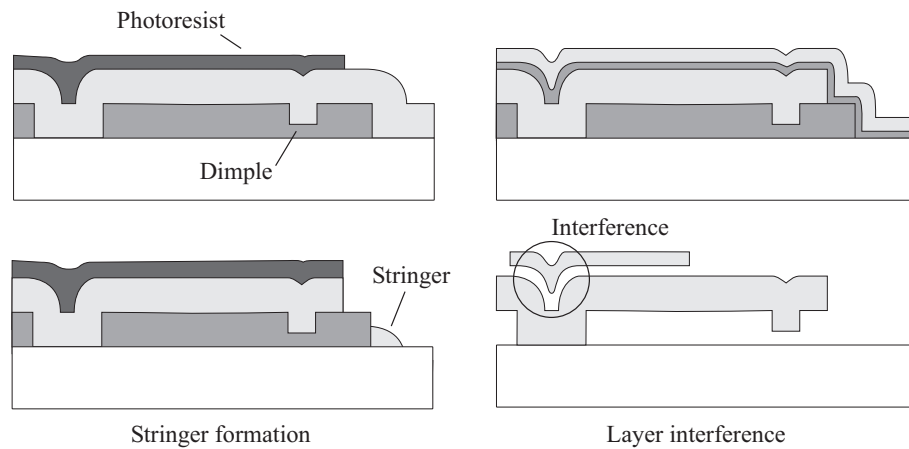


Figure 3.32: Common surface micromachining issues.

A common problem is the formation of strings of structural material, called ‘stringers’, during the patterning of the upper layer. Actually the high anisotropy of the etching by RIE leaves some material where the layer is thicker because of the conformal deposition of the structural material. To avoid the problem during fabrication, the RIE etching time needs to be substantially increased to fully etch the layer where it is thicker. For example the MUMPS surface micromachining process proposed by the foundry Memscap is using an overetching of 100%, that is, the etching lasts twice the time needed to clear the material in the flat zone. Another common issue is the likelihood of structure interference between the stacked layers. In Figure 3.32 we see that the topography creates an unintended protrusion below the top structural layer that will forbid it to move freely sideways - probably dooming the whole device. This problem can be tackled during layout, particularly when the layout editor has a cross-section view, like L-Edit from Tanner Research. However even a clever layout won’t be able to suppress this problem completely and it will need to be addressed during fabrication. Actually it is possible to polish using Chemical-Mechanical Polishing (CMP) the intermediate sacrificial layer to make it completely flat, will avoid all interference problems. For example, Sandia National Laboratory uses oxide CMP of the second sacrificial layer for their four layers SUMMiT V process.

However, sometimes the interference may be a desired effect and for example the so called ‘scissors’ hinge [26] design shown in Figure 3.33 benefits greatly from it. The scissors hinge is designed to provide a hinge functionality with micromachining process and as we see here the protrusions below the upper layer help to hold the hinge axis tightly. If we had to rely on lithography only, the gap between the axis and the fixed part in the first structural layer would be at best  $2\ \mu\text{m}$ , as limited by the design rules, and the axis will have too much play. However the

protrusions below the staple reduce the gap to  $0.75\ \mu\text{m}$ , the thickness of the second sacrificial layer, and the quality of the hinge is greatly increased.

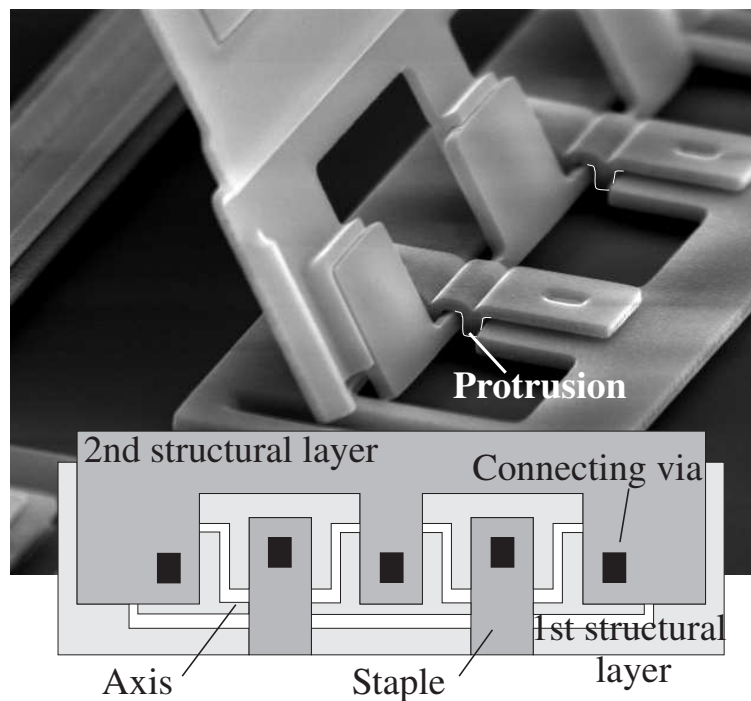


Figure 3.33: Tight clearance obtained by layer interference in a hinge structure.

The final step in surface micromachining process is the release - and this critical step has also a fair amount of issues that need to be considered.

### 3.5.3 Microstructure release

The release step is the source of much technologist woes. Release is usually a wet process that is used to dissolve the sacrificial material under the structure to be freed. However the removal rate is usually relatively slow because the sacrificial layer is only a few  $\mu\text{m}$  thick and the reaction becomes quickly diffusion limited. Then the depth of sacrificial layer dissolved under the structure will increase slowly with the etching time as

$$d_{\text{release}} \propto \sqrt{t_{\text{etch}}}.$$

Simply said, releasing a structure twice as wide will take 4 times more time. However if the etching lasts too long the chemical may start attacking the device structural material too. A first measure to avoid problems is to use compatible material and chemical, where the sacrificial layer is etched quickly but other material not at all. A typical example is given by the DLP (Digital Light Processing) from Texas Instruments, where the structural layer is aluminum and the sacrificial layer is a polymer. The polymer is removed with oxygen plasma, and prolonged

release time will only slightly affect the metal.

This ideal case is often difficult to reach and for example metals have often a finite etch rate in HF, which is used to remove PSG sacrificial layer. Thus to decrease the release time we have to facilitate etching of the sacrificial layer by providing access hole for the chemical through the structural layer. In the case of Figure 3.25 for example, the mirror metal starts to peel off after about 10 minutes in HF. However in about 5 minutes HF could only reach  $40\ \mu\text{m}$  under a plain plate, and the designer introduced 'release holes'. These holes in the structural layer are spaced by roughly  $30\ \mu\text{m}$  in the middle of the mirror plate (the white dots in the figure) allowing for the HF to etch all the oxide beneath in less than 5 minutes.

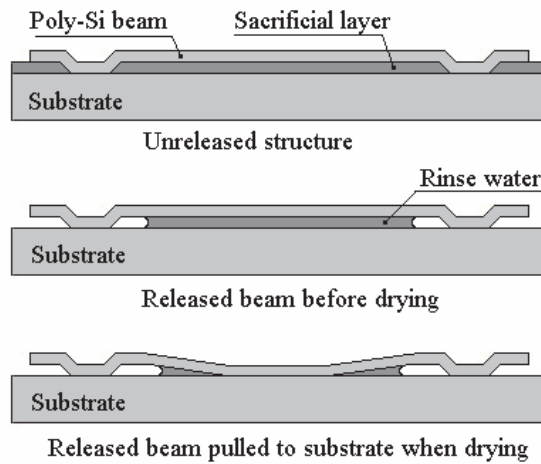
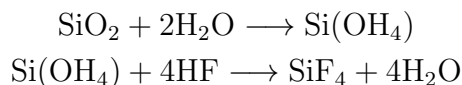


Figure 3.34: Stiction phenomenon during release.

The problems with wet release continue when you need to dry your sample. The meniscus created by the receding liquid/air interface tends to pull the structure against the substrate. This intimate contact give rise to other surface forces like Van der Waals force, which will irremediably pin your structure to the substrate when the drying is complete, effectively destroying your device. This phenomenon is referred as stiction (Figure 3.34). Strategies that have been used to overcome this problem have tackled it at design and fabrication level. In surface micromachining the idea has been to reduce the contact surface by introducing dimples under the structure. From the fabrication side, super-critical drying, where the liquid changes to gas without creating a receding meniscus, has also been applied successfully. Coating the structure with non-sticking layer (fluorocarbon, hydrophobic SAM...) has also proved successful and this method, albeit more complex, has the added advantage to provide long lasting protection again sticking that could arise during use.

Finally, a completely different approach is to avoid wet release altogether and instead to perform a dry release with a gas or a vapour, suppressing entirely the stiction concerns. For example the Multi-Project-Wafer (MPW) process run by

Bosch uses HF vapour to remove the oxide layer below the polycrystalline structures. The reaction is roughly as follow:



where all the final by-products are, of course, gaseous. The main issue with the technique is the high toxicity of the HF vapour and in Table 3.8 we describe two other popular methods which present less risk: dissolving polymer sacrificial layer with  $\text{O}_2$  plasma, and using xenon difluoride ( $\text{XeF}_2$ ) to etch sacrificial silicon. The xenon difluoride is a gas showing an excellent selectivity, having etching rate ratio close to 1000 with metal and up to 10000 with oxide. The gas has thus been used successfully to release very compliant or nano-sized oxide structures where silicon was used as the sacrificial material. The process does not use plasma, making the chamber rather simple, and several manufacturers like XactiX (in cooperation with STS), in the USA or PentaVacuum in Singapore are proposing tools exploiting the technology.

### 3.6 DRIE micromachining

Deep reactive ion etching (DRIE) micromachining shares features both from surface and bulk micromachining. As in bulk micromachining the structure is etched in the bulk of the substrate, but as in surface micromachining a release step is used to free the microstructure. Figure 3.35 shows a simplified process of bulk

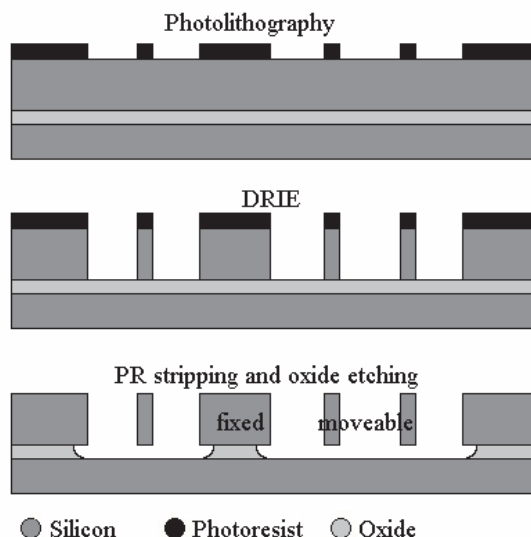


Figure 3.35: Bulk micromachining of SOI wafer by DRIE.

micromachining on silicon-on-oxide (SOI) wafer using deep reactive ion etching

(DRIE), a special MEMS dry etch technique allowing large etch depth with very vertical side walls. The SOI wafers used in MEMS usually have a device layer thickness between 10 and 200 $\mu\text{m}$  where the structure is defined. After photolithography, the wafer is etched with DRIE to form high aspect ratio silicon structures, and the buried silicon dioxide acts as an effective etch stop. Stripping off the protective photoresist by  $\text{O}_2$  plasma and then etching the sacrificial layer of the oxide using HF to release the microstructure finish the device. This simple, yet powerful, technique needs only one mask to obtain working devices, and it is understandably used in commercial products. The best known example is the optical switch produced by Sercalo, a company founded by C. Marxer the inventor of this fabrication technique.

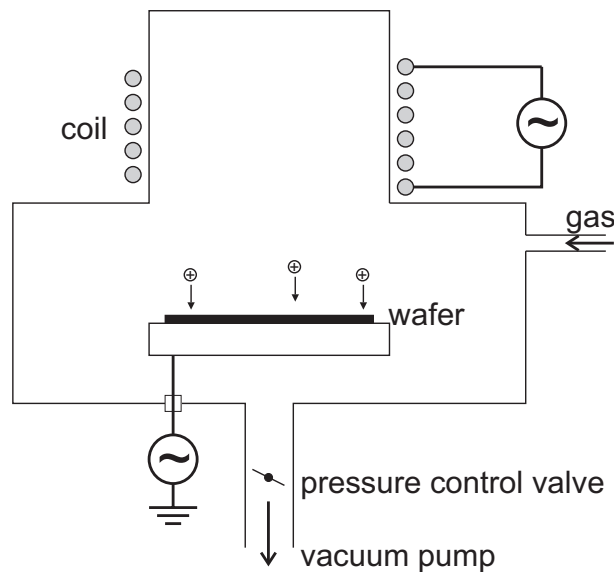


Figure 3.36: ICP-DRIE system.

DRIE has reached a large popularity in recent years among MEMS community and the tools produced by Adixen (Alcatel), Surface Technology Systems (STS) and Oxford System produce high aspect ratio structures ( $>25$ ) with vertical side-walls ( $>89^\circ$ ) at a decent etch rate (6 $\mu\text{m}/\text{min}$  or more). A standard DRIE system, as shown in Figure 3.36, uses high density inductively coupled plasma (ICP) as the plasma source. In the ICP plasma source the electrons are set in motion by a RF magnetic field, creating plasma with much higher density than with capacitively coupled source. Additionally, a second RF source is placed on the wafer to increase the ion bombardment on the wafer by polarizing it negatively with respect to the plasma. By using the two sources we can decouple the ion generation and energy, thus allowing a much finer control of the etching condition. We can get a lot of energy in the ICP (e.g. 3 kW) to generate large ion density and high etching speed simultaneously with a mild accelerating voltage (e.g. 100 W) on the wafer to remain in the RIE regime and avoid mechanically chirping matter out of the

wafer.

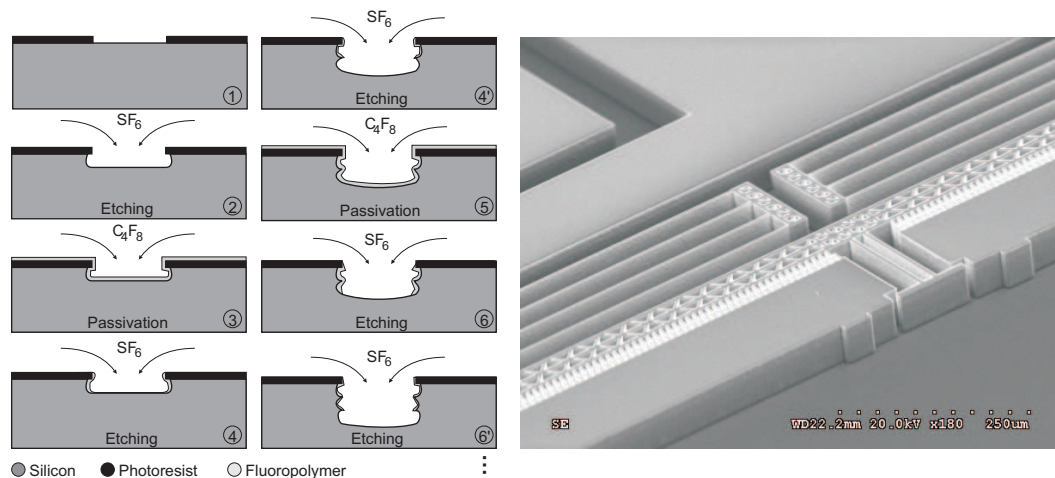


Figure 3.37: Principle of the Bosch process for DRIE etching and 50  $\mu\text{m}$  thick movable mirror fabricated on SOI wafer.

In addition to increase the anisotropy for deep etch, DRIE usually adopts the patented “Bosch process” (Figure 3.37). The Bosch process is a repetition of two alternating steps: passivation and etching. In the passivation step,  $\text{C}_4\text{F}_8$  gas flows into the ICP chamber forming a polymer protective layer ( $\text{n}(-\text{CF}_2-)$ ) on all the surfaces. In the following etch step, the  $\text{SF}_6$  gas in the plasma chamber is dissociated to F-radicals and ions. The vertical ion bombardment sputter away the polymer at the trench bottom, while keeping the sidewall untouched and still protected by the polymer. Then the radicals chemically etch the silicon on the bottom making the trench deeper. By carefully controlling the duration of the etching and passivation steps, trenches with aspect ratio of 25:1 are routinely fabricated - and aspect ratio as high as 100:1 have been reached. Figure 3.37 right shows a SEM picture of a movable mirror fabricated by DRIE on a SOI wafer.

The DRIE is a very versatile tool and allows a good control on the etched profile slope, mostly by varying the etching/passivation duration and also by varying the substrate biasing voltage. However, it is affected by common RIE problems like microloading, where etching rate is slower for high density patterns than for low density ones. This effect is linked with the transport speed of reactant and products to and from the surface, which can be improved somewhat by increasing the flow rate of etching gas. But this is not the only issue, and other common issues are shown in Figure 3.38.

One issue with DRIE is the presence of regular ripple with an amplitude over 100 nm on the vertical edge of trenches, a phenomena referred to as scalloping. The ripples comes from the repetition of isotropic etching and passivating steps, and can be annoying for etching nano-pillars with a few 100 nm diameter or for obtaining vertical wall with smooth mirror finish. Actually, they can be mostly removed by shortening the etching step to 1 s, instead of a standard 7 s, and

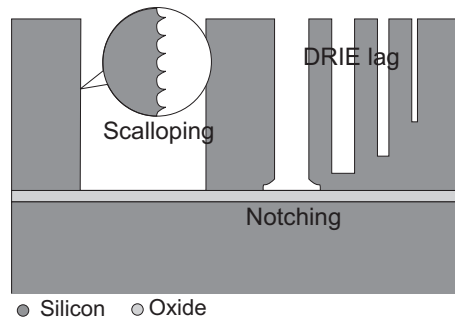


Figure 3.38: Some issues affecting DRIE process: scalloping, notching and lag (ARDE).

by reducing the passivation step duration accordingly. This of course results in a much slower etching rate, but is a usual practice of trading etching speed for improving another etching parameter.

The existence of DRIE lag is also a nuisance that needs to be considered. Actually, in narrow trenches, ion charging at the sidewall lowers the electric field and the energy of the ions, decreasing etching rate as compared to what happens in wide trenches. This is described as the aspect-ratio dependent etching (ARDE) effect. This effect again can be controlled by properly tweaking the recipe and trading a bit of etching speed.

Another major issue existing in DRIE is the fast silicon undertech that happens in SOI trenches when the etch reaches the buried oxide layer. Actually after the silicon has been completely etched, the oxide get exposed, and positive charges build-up in the insulating layer. This local space charge deviates the incoming ions laterally, causing an increased etch at the lower portion of the sidewall of the trench, an effect called notching. The most recent DRIE tools have managed to tackle this problem satisfactorily, by using a low frequency biasing scheme. Actually the normal RIE plasma frequency (13.56 MHz) is too high to have any effect on the ions, but by lowering the frequency to 380 kHz the ions bombardment will follow the field. In a way similar to what happens in a RF sputter, but this time for the ions, the ions during the positive bias pulse won't be anymore directed toward the substrate. The plasma electrons will then be attracted there and recombine within the charged insulator, suppressing the spatial charge. By varying the cyclic ratio of the low frequency bias pulse it is thus possible to control the etching/uncharging timing, and obtain optimal etching rate while avoiding notching.

It should be noted that the notching effect can be put to good use and help produce an 'etch and release' process. Actually it has been found [24] that the notching effect is self limiting and the depth of the notch is roughly equal to the width of the trench as soon as the trench has an aspect ratio larger than 2 (for smaller aspect ratio there is no notching effect). In this way, by carefully designing the geometry of the layout, it is possible to etch the structure and finally obtain anchored or free structures within the same etching step. This simplifies further the DRIE



fabrication process, and the device can now be operated right after emerging from the DRIE - without the need for a separate release etch!

The SOI wafer used often in DRIE machining is still expensive and it is possible to obtain the thick silicon structural layer by growing it using epitaxy on an oxidized wafer. Even more simply, DRIE has been used to etch through the Si wafer for a dry etched version of bulk micromachining but allowing complete freedom over layout as there is no more crystallographic orientation concerns. In this case wafer bonding can be used to provide movable part.

## 3.7 Other microfabrication techniques

### 3.7.1 Micro-molding and LIGA

Other methods exist for patterning where no material is removed but where it is simply molded. LIGA, a German acronym for lithography (Lithographie), electroforming (Galvanoformung), and molding (Abformung) is the mother of these methods. LIGA makes very high aspect ratio 3-D microstructures with non-silicon

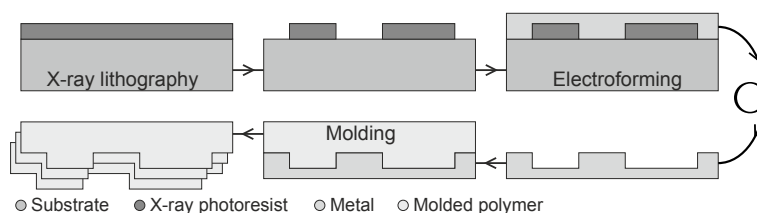


Figure 3.39: The LIGA process.

materials such as metal, plastic or ceramics using replication or molding. The general view of the LIGA process is shown in Figure 3.39. The process begins with X-ray lithography using a synchrotron source (e.g. energy of 2.4 GeV and wavelength of 2 Å) to expose a thick layer of X-ray photoresist (e.g. PMMA). Because of the incredibly small wavelength, diffraction effects are minimized and thick layer of photoresist can be patterned with sub-micron accuracy. The resist mold is subsequently used for electroforming and metal (e.g. nickel using  $\text{NiCl}_2$  solution) is electroplated in the resist mold. After the resist is dissolved, the metal structure remains. This structure may be the final product but to lower down the costs, it usually serves as a mold insert for injection molding or hot embossing. The possibility to replicate hundreds of part with the same insert opens the door to cheap mass production.

When the sub-micrometer resolution is not much of a concern, pseudo-LIGA processes can be advantageously used. These techniques avoid using the high cost X-ray source for the mold fabrication by replacing it by the thick photoresist SU8 and a standard UV exposure or even by fabricating a silicon mold using DRIE.

### 3.7.2 Polymer MEMS

Bulk and surface micromachining can be classified as direct etch method, where the device pattern is obtained by removing material from the substrate or from deposited layers. However, etching necessitates the use of lithography, which already includes patterning the photoresist, then why would we want to etch the lower layer when the pattern is already here? Actually lithography for MEMS has seen the emergence of ultra-thick photoresist that can be spun up to several 100  $\mu\text{m}$  and exposed with a standard mask aligner, providing a quick way to the production of micro-parts. SU8, a high-density negative photoresist can be spun in excess of 200  $\mu\text{m}$  and allows the fabrication of mechanical parts [25] of good quality. It is used in many applications ranging from bioMEMS with micro-parts for tissue scaffold or channels, for example to packaging, where it is used as buffer layer.

Another application of thick photo-patternable polymer is the fabrication of mi-

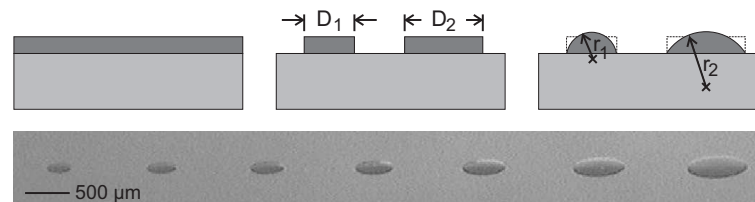


Figure 3.40: Fabrication of microlenses using reflow of polymer.

cro-lenses as shown in Figure 3.40. After patterning pillars of appropriate diameter in photoresist, the key process step is to place the wafer at a temperature higher than the glass transition temperature of the polymer (e.g. 115°C for AZ9260 photoresist). The polymer will melt, and due to surface tension forces, will assume a spherical shape – the classical shape of the lens. The interest of this technique is that the continuous profile of the lens, which would have been hard to obtain using etching method, is obtained here through a fundamental principle of nature, the minimization of energy in a system, which translates itself in the minimization of surface energy at this scale. Varying the diameter of the pillar before this so called reflow process allows obtaining different radius of curvature, that is, different focal length. Another option would be to change the thickness of the photoresist layer, as the final shape is mostly determined by the volume of photoresist in the original pillar. One of the interest of this technology is that polymers usually have better optical properties than silicon in the visible, and there is a lot of opportunity for polymer micro-optical elements and system, a domain that is sometimes called Polymer Optical MEMS or POEMS.

Next to these major techniques, other microfabrication processes exist and keep emerging. They all have their purpose and advantages and are often used for a specific application. For example, quartz micromachining is based on anisotropic wet etching of quartz wafers to take benefit of its stable piezoelectric properties

and build sensors like gyroscopes.

## 3.8 Characterization

The small dimension of the MEMS makes it hard to properly measure their geometry or observe their operation by simply using rulers or our naked eyes. Accordingly, a large range of specialized tools including some specifically developed for the MEMS industry, are used for interfacing with the micro-world and measure geometry, layer thickness, beam motion, materials properties, etc.

We list in Table 3.10 some of the most common measurements tools for the measurands encountered in MEMS. We note that most measurand can usually be obtained with different tools, but the tool will usually differ in other characteristics, as whether it is an area or a point measurement, or a contact or a non-contact method, etc. For example, measuring surface roughness may be obtained with a stylus profilometer, which is a contact method working point by point or with an optical interferometer, which is non contact method and records a complete surface simultaneously. For a proper choice of the right instrument, additional properties will often need to be considered and, for example, the optical interferometer will have difficulty to work with transparent samples and will usually have a smaller range than a stylus profilometer. Clearly, the ability to work with multiple tools is important for answering all the challenges of MEMS measurement.

In the following sections we will describe a few of these tools in more details, but be convinced that good characterization skills will only be acquired with a knowledge of the capabilities of more tools than what is cited in the table.

### 3.8.1 Light Microscope

The light microscope is ubiquitous in MEMS characterization, letting the micro-world come to our sight. It is used repeatedly in the cleanroom at the end of each process steps for quality control or after fabrication for observing the operation of the completed MEMS. From the light microscope we have been using for biological sample viewing in 3rd grade to the fluorescent confocal microscope there is a complete palette of microscopes available for different usage, and for MEMS characterization the most used microscope is called a reflected light infinity corrected compound microscope (Figure 3.41). The compound part just means it uses two sets of lenses for magnifying the sample, the infinity corrected is a cool thing we will explain latter while the reflected part (as opposed to transmitted light used for biologic sample microscopy) means that the optical path for observation and illumination are from the top allowing opaque samples observation – which are the norm in MEMS. Such microscope is a very precise optical component, requiring high tolerance manufacturing and careful design to obtain all the desired features for precise characterization.

Type	Mesurand	Tool	Remark
Geometry	in-plane	optical microscope, SEM	
	depth	stylus profilometer, AFM	contact
		optical mic., interferometer SEM (SE)	non-contact destructive
Physical	thickness	ellipsometer	dielectric
		IR reflectometry	metals
	roughness	stylus profilometer, AFM	contact
		interferometer	non-contact
		refractive index	ellipsometer
	resistivity	four-probe measurement	
	surface energy	sessile drop, pendant drop	
	interfacial tension	Noüy ring, Pt plate	
	modulus	nano-indenter, instron	
	Chemical	composition	SEM (EDS), SEM (BSE)
XPS, SIMS, AES, NMR			
structure		XRD TEM	
Dynamics	vibration	LDV	point meas.
		strobed interferometer	area meas.

Table 3.10: MEMS characterizatton techniques and tools.

Although its design details are complex the principle of the compound microscope is simple : the objective forms an enlarged and inverted image of the object at 160 mm (normalized length) from the objective end and it is this real image that is observed with the ocular (also called the eyepiece) allowing further enlargement. The imaging light path is shown in Figure 3.42, where we have used the principal planes to represent the equivalent optical system for each set of thick lenses<sup>14</sup>. Actually we represent in the figure an infinity corrected microscope where

<sup>14</sup>The equivalent ray path shown in the figure (using principal planes) is useful for obtaining

the objective lens is used to form an image “at infinity” (instead of the standard 160 mm) thus requiring that the object is placed in the focal plane of the objective ( $f_o$ ). Then we use an additional lens (the tube lens) placed after the objective to form in its focal plane ( $f'_T$ ) the real image of the object. The space between the objective and the tube lens provides a path of parallel light where optical components (like polarizers, beam-splitters...) can be inserted or removed without changing the optical path, maintaining the same observed position on the sample. Finally the eyepiece object focal plane ( $f_E$ ) is positioned at the intermediate image plane for forming an image at infinity. In this case the eye does not need accommodation (that is, tense or relax the eye lens) and observation will remain comfortable for long duration.

For MEMS observation, the main parameters of interest in the microscope are:

---

the actual position and size of the image but between the object and image conjugate points “real” rays would follow a different path and propagate by bending at each lens interface according to Snell-Descartes law.

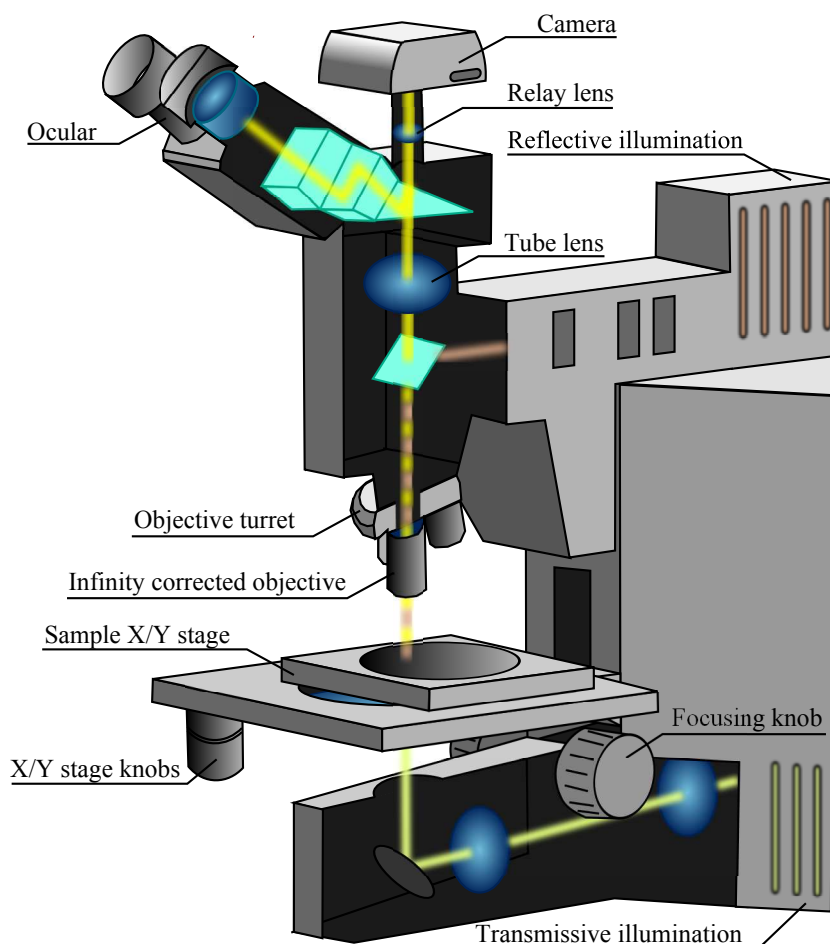


Figure 3.41: Compound microscope for reflected and transmitted light observation.

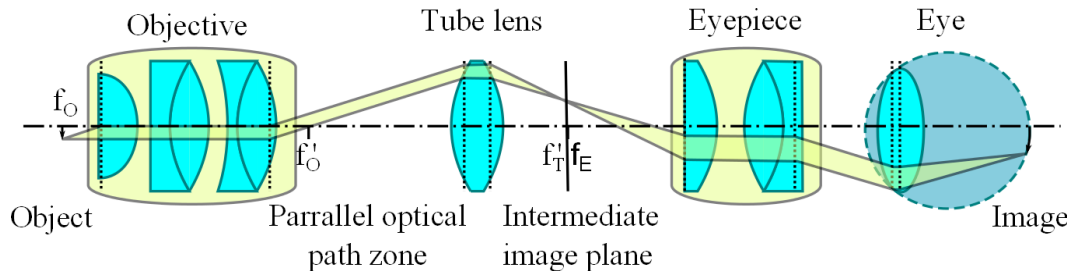


Figure 3.42: Formation of magnified image in a compound microscope with infinity corrected objective (ray propagation is shown based on principal planes, not lens interface).

the magnification, the resolution, the depth of field and the working distance.

**Magnification** is the parameter describing the apparent enlargement of the object when it is observed through the microscope. It may not be completely obvious, but the concept describes different things if the image is observed with the eye through the ocular or observed with a camera.

For an observation through the eyepiece, the magnification is the product of two terms  $M = M_O M_E$ : one term due to the objective  $M_O$  and one due to the eyepiece  $M_E$ . As the objective (or the combination of the objective and the tube lens in case of infinitely corrected objective) is producing a real image in the intermediate plane, the magnification is simply computed as the ratio between the size of the object and the size of its image (the so called lateral magnification). As the angular dimension of the object (i.e., the angle subtended by the object) and of its image are the same (this property is directly linked to Snell-Descartes law), it is simple to show that the lateral magnification of the objective is actually equals to  $M_O = 160 \cdot 10^{-3} / f_O$  the ratio of the standard distance<sup>15</sup> from the objective principal plane to the intermediate image plane (called the tube length) and the objective focal length (in the case of infinitely corrected objective we have  $M_O = f_T / f_O$  as shown in Figure 3.42). The magnification of the eyepiece is computed differently as it is producing a virtual image like a simple magnifier. In this case we use the concept of angular magnification, the ratio between the angle that the object subtends if it is placed at 25 cm<sup>16</sup> from the eye and the actual angle subtended by the object when it is magnified. We also note that the angular dimension of the object is directly linked to the dimension of its image on the retina, which is given for parallel rays entering the eyes by  $\delta l = 22 \delta \alpha \text{ mm}$  where 22 mm is the focal length of the eye. Again it is simple to show that this definition is equivalent to saying that  $M_E = 250 \text{ mm} / f_E$ .

<sup>15</sup>A few manufacturers do not use this standard length of 160 mm.

<sup>16</sup>This reference distance of 25 cm corresponds to the minimum distance where “normal” unaided eye can focus.

Finally, the angular magnification of the microscope can be given by:

$$M = \frac{160 \cdot 10^{-3}}{f_O} \frac{250 \cdot 10^{-3}}{f_E} \text{ for classical microscope} \quad (3.6)$$

$$M = \frac{f_T}{f_O} \frac{250 \cdot 10^{-3}}{f_E} \text{ for infinitely corrected microscope} \quad (3.7)$$

We note that the objective is inverting the image and the eyepiece isn't changing orientation (magnification is positive) thus the final image should be inverted. However, the prisms placed before the eyepiece that can be seen in Figure 3.41 ensure by using folded reflection that the actual image is erect. The magnification is changed by rotating the turret and selecting an objective with a higher or lower magnification. Interestingly, although we change the focal length (hence their magnification), the objective are designed so that even for objective corrected for finite tube length, the position of the intermediate image does not change, a fact known as parfocality. In practice, there is little reference to focal length on objectives and oculars and the lateral magnification is directly labeled on the objective (common value are  $2\times$ ,  $5\times$ ,  $10\times$ ,  $20\times$ ,  $40\times$ ,  $100\times$ ) while the angular magnification of the eyepiece is commonly written as  $10\times$  or  $20\times$ .

When we use a camera we need to form a real image on its photosensitive sensor, thus we can not use the ocular (it produces a virtual image) but should place the sensor directly after the objective. The issue with this simple configuration is that the camera can not be placed in the intermediate image plane inside the microscope (Figure 3.41) but can only be placed further away "outside" of the microscope. Accordingly the object would have to be moved using the focusing knob to change the position of the intermediate image and focus the image through the objective directly on the sensor. However this is not very practical because if we switch back to visual observation through the eyepiece the intermediate image won't be in focus and the focusing knob would have to be used again<sup>17</sup>. The modern microscope solves this problem by using an additional relay lens (Figure 3.41) forming a real image of the intermediate image on the camera sensor. This lens is normally placed so that the magnified intermediate image fills the camera sensor and, depending on its format the magnification is normally between  $0.5\times$  and  $1.25\times$ . The magnification of the microscope is then simply equal to the lateral magnification of the objective multiplied by the relay lens magnification. Note that because we use two lenses with negative magnification, the resulting image will be erect. Actually here we only considered the optical magnification, but when we use a camera, we may also want to consider the electronic magnification: for optimal magnification,

---

<sup>17</sup>Additionally, and more importantly, the objective will correct optical aberrations for a particular distance between the object and the image – change this distance and some aberration will crop up again.

a pixel on the image will be a magnified view of one sensitive element on the camera. For visual observation this factor can be obtained by simply taking the ratio of the image dimension on the screen with the camera sensor dimension. For example, for an image that comes from a camera with a 1" sensor (with an actual diagonal of 16 mm) filling completely a 17" (or 43 cm) screen, the electronic magnification is  $430/16 \approx 27$ . Note that the format of the sensors labeled as 1/3", 1/2", 2/3" or 1", does not directly translate to actual dimension and we would need to refer to Table 3.11 The total

Format	Diagonal length	Length	Height
1/3"	6	4.8	3.6
1/2"	8	6.4	4.8
2/3"	11	8.8	6.6
1"	16	12.8	9.6

Table 3.11: Camera sensor size.

visual magnification will be the product of the optical magnification and the electronic magnification. In general, as the exact magnification factor is not easy to come by, the software used for microscope image processing will allow using a calibration procedure. In this case an object of known dimension is first placed under the objective and measured (in pixels), then the software will apply the same scaling factor in  $\mu\text{m}/\text{pixel}$  to all the images recorded with the same objective. Normally when we change objective, the scaling is just multiplied by the ratio between the reference objective magnification and the current objective magnification, as the electronic magnification will remain the same, but some slight adjustment may require a new calibration.

**Resolution** is also called the resolving power of the microscope and gives the shortest distance between two points on a specimen that can still be resolved. The resolution depends ultimately on physics, and more specifically on diffraction law.

We note that the resolving power does not describes the ability to observe small structure, but the capability to resolve between two neighboring points or observe detailed features. The difference is of importance, and, for example, nanoparticle below 50 nm that disperse enough light can still be "observed" (that is, we see there is a halo, but can not really tell its details and only guess that the particle stays at the center of the halo) with a light microscope.

The possibility to correct the optical aberration of objectives has tremendously improved since the apparition of simulation software. Nowadays, most optical system are mostly limited by the diffraction effect and no more



by spherical or other aberration. Diffraction by the microscope objective acting as the entrance pupil of the microscope can be simply considered as a problem of diffraction by a circular aperture in the far field - with infinity corrected objective we observe “at infinity” - and we can use the simplified model of Fraunhofer diffraction (cf. Appendix F)<sup>18</sup>. This means that the image of a point object observed by the objective and formed by the tube lens in an infinitely corrected microscope is not a point, but an Airy disk. Then we may define criterion for resolution, based on whether the Airy disk produced by two neighbouring points can be distinguished or not. The meaning of “distinguished” is subject to debate and one of the earliest criterion has been postulated by Lord Rayleigh: two points of the same size can be distinguished if the center of the two resulting Airy disks is separated by a distance equal to the diameter of the Airy disk (Figure 3.43-c). Another way to look at this criterion, and what makes its simplicity, is to say that the maximum of the Airy disk should be located at the first zero of the other Airy pattern. According to Rayleigh’s criterion we may obtain the resolving

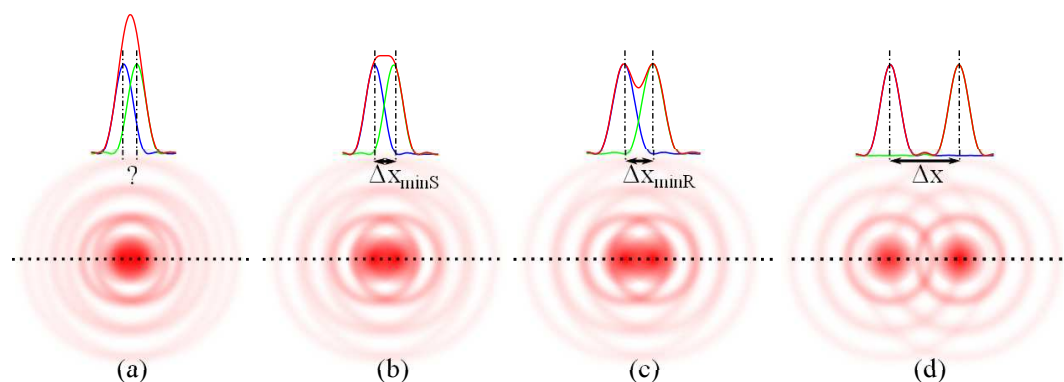


Figure 3.43: Resolving two neighbouring punctual objects having the same Airy disk (a) not resolved (b) barely resolved according to Sparrow’s rule (c) barely resolved according to Rayleigh’s rule (d) resolved.

power of the objective using the formulas derived in Appendix F as :

$$\delta x_{minR} = 1.22\lambda \frac{f_O}{D}$$

where  $D = 2a$  is the diameter of the objective entrance pupil and  $f_O$  its focal length.

The Rayleigh’s criterion is not the only one possible and other, less stringent ones, have been proposed like the Sparrow’s rule which is particularly representative of what is actually observed with a telescope. Here the idea is to

<sup>18</sup>Actually it is an approximation as the far-field hypothesis is not entirely verified here: the observation plane is indeed far but the wave from the very close object striking the objective aperture is not plane.

say that two points can not be resolved anymore when the dip between the two maximums disappears and we get a flat topped signal (Figure 3.43-b). In this case the limit is slightly improved and we get :

$$\Delta x_{minS} = 0.96\lambda \frac{f_O}{D}$$

Of course the approximation we made for modeling the objective diffraction (remember we don't really have Fraunhofer diffraction after all), does not make the 25% improvement really meaningful, and a rule of thumb that the resolution is of the order of

$$\Delta x_{min} \approx \lambda \frac{f_O}{D}$$

is more than enough.

It is interesting to note that, in a way, if we include the eye in the system, the overall resolution rapidly stops increasing with magnification. Actually, the angular resolution limit of the eye is given in many document as 1' angle and its focal length at about 22 mm, corresponding to an image of roughly 6  $\mu\text{m}$  on the retina – about the size of the light sensitive elements there. Thus for an objective resolving power of about 300 nm we already get the 6  $\mu\text{m}$  limit with a mean 30 $\times$  magnification ! The actual resolution of a 'normal' eye may be 2 to 3 times lower (2' or 3'), thus we may expect that in reality magnification up to 100 $\times$  still induce an improvement. Still, we should not deduce that larger magnification is useless (it is relatively easy to obtain with a microscope 500 $\times$  or 1000 $\times$ ) because a larger magnification will allow more comfortable observation, and the Airy disk of the points observed through the microscope will be imaged on a group of eye sensitive elements, getting an improved resolution by allowing to resolve the detail of this disk. The same things happens with a camera. As soon as the magnification produces an Airy disk of the objective matching with the size of the camera pixel, a further increase in magnification will only increase the resolution by allowing sub-sampling of the Airy disk profile.

**Depth of field** is the number describing the *longitudinal* resolving power (axial resolution along the focus direction) as opposed to the *lateral* resolving power, just described. Actually, for an infinity corrected microscope, a flat object placed in the objective focal plane  $f_O$  will be perfectly imaged in the intermediate image plane. However if the object is offset and placed slightly above or below  $f_O$ , its image won't be right at the intermediate image plane but slightly before or after. Thus in the intermediate image plane which is observed through the ocular (or by the relay lens), this last object would appear slightly blurred. The depth of field gives an estimate of how much offset can be tolerated before the image is too blurred.

This effect has two main contributors (neglecting again aberrations) a geometrical effect and the diffraction due to wave optics. The geometrical effect can be obtained by first observing that for an object on the optical axis there is a region symmetrical around the image plane where the image of a punctual object will give a disk with a diameter smaller than  $\delta$ . If  $\delta$  is small (on the order of the diffraction effect), this geometrical blur can be ignored and for practical purpose we would consider that the object is “on focus”. Then, the depth of field (DOF) is actually the range where the object can be placed in front of the objective and still yield an image in this zone of acceptable blur (Figure 3.44). We note that this region is not symmetrical around the nominal focal plane in front of the objective but is longer “before” it than “after”. The diagram in the figure allows after some calculation and sim-

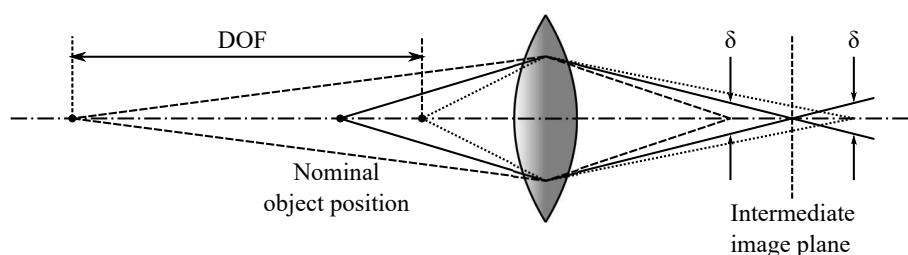


Figure 3.44: Depth of field due to geometrical effect.

plification to obtain a relatively simple expression for the DOF. Similarly, the wave optics effect can be obtained by considering the diffraction in more details and after combining the two expressions obtained for each source we find :

$$DOF = n\lambda \left( \frac{D}{2f_0} \right)^2 + \frac{n\delta}{M_O} \frac{D}{2f_0}$$

where  $n$  is the index of refraction in front of the objective and  $M_O$  the lateral magnification of the microscope objective. For high  $D/f_0$  ratio (and magnification  $M_O$ ) of the microscope, depth of field is determined primarily by wave optics (first term of the equation), while at lower value of  $D/f_0$  ratio, the geometrical effect (second term) dominates the phenomenon.

**Numerical Aperture** is a critical value that indicates the light gathering power of the objective. The numerical aperture (Figure 3.45) is defined by :

$$NA = n \sin \alpha$$

where  $n$  is the index of refraction of the medium at the lens entrance and  $\alpha$  the incidence angle of the most extreme ray that will be accepted by the entrance lens without hitting a stop somewhere down the optical system. Interestingly we note that from Snell-Descartes law, this quantity is invariant across all interfaces in the optical system, thus it won't change if a parallel

side glass plate is placed in front of the objective or if the objective is immersed in a liquid. For a microscope objective, the numerical aperture is directly

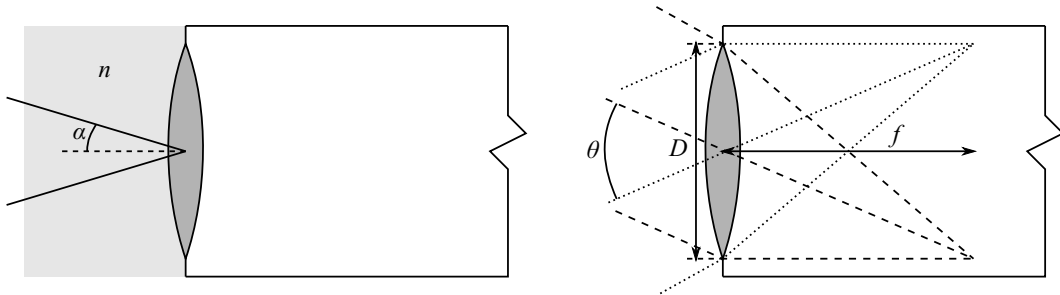


Figure 3.45: Numerical aperture NA and relative aperture  $f/\#$ .

related to a quantity known to photographers: the relative aperture, also called the f-number  $f/\#$ . This second variable also reveals the light gathering capability of an optical system by saying that this ability is proportional to the entrance aperture (pupil) surface, thus proportional to  $D^2$ , and inversely proportional to the size of the image in the focal plane, thus proportional to  $1/f^2$ . Then the light gathering power is proportional to  $D^2/f^2 = (D/f)^2$ . In practice people had more use for the light “stopping” property, that is the opposite of the light gathering ability, and for describing this property we thus use the quantity  $f/D$ , the relative aperture. For example a lens of aperture 3 mm with a focal length of 6 mm will have a relative aperture of  $f/D = 6/3 = 2$ , which is curiously noted  $f/2$  - the f-number. Actually photographers know too well that when one increases the aperture by one stop one is dividing the quantity of light hitting the film or the sensor by a factor of 2 – actually the standard series of f-number for camera objective diaphragm is 1, 1.4, 2, 2.8, 4, 5.6, etc and we have a ratio of  $\sqrt{2}$  between two consecutive stops in the series, explaining the factor of  $2 = \sqrt{2}^2$ .

In the case of microscope objective it is striking to see the close relationship between  $f/\#$  and NA. Actually by looking at Figure 3.45, we see that for small angle we have  $\frac{D}{f} \approx \theta$  thus  $f/\# \approx 1/\theta$ . In the case of  $n = 1$  (objective is placed in air) we have  $NA \approx \sin \alpha \approx \alpha$  the latter approximation being valid for small angle. As  $\theta = 2\alpha$  is the solid angle of the cone of accepted rays in the optical system, we see that  $f/\# \approx 1/(2NA)$ , the f-number is inversely proportional to NA.

What makes NA interesting is that it non only directly determines the light gathering power but also the resolving power, and the depth of focus of the objective. By replacing in the previously derived equations  $D/2f_O$  by NA we find the dependency shown in Table 3.12. That means, for example, that the larger the numerical aperture of the objective is, the smaller its depth of field and the larger its resolving power will be.

Parameter	Expression	Improves with
Light gathering		$NA^2$
Resolving power	$\Delta x_{min} \approx \frac{\lambda}{2NA}$	NA
Depth of field (diffraction)	$\frac{n\lambda}{NA^2}$	$1/NA^2$

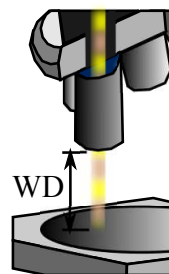
Table 3.12: Influence of numerical aperture on main microscope characteristics.

**Working distance** is the distance that is existing between the end of the microscope objective and the sample as shown in the inset.

In general this distance decreases rapidly with the focal length and the numerical aperture of the objective, and if standard  $10\times$  objective have a comfortable working distance of about 10 mm, the  $100\times$  can only boast a meager 0.5 mm or less. This short distance poses problem for MEMS characterization as the probe used to power the device will easily interfere with the microscope objective.

Luckily, for infinity corrected objective this distance is normally larger and the manufacturers have developed special long working distance objective (LWD) particularly useful for MEMS application.

However, the objective still requires a large NA for maintaining the high resolution and this conundrum can only be solved by using large lens for the objective input lens, upping their price by a factor of 10 when compared to classical objectives. Some brand like Mitutoyo are proposing  $10\times$  objective with a working distance of 33 mm and  $100\times$  exceeding 6 mm while keeping  $NA = 0.7$ .



As we see on the inset, many information are written on the objective barrel: the type of objective, usually reflecting the correction of chromatic aberration (best is Aplanat corrected for 4 wavelengths, less good is Achromat corrected for 2 wavelengths) and of field-curvature aberration (Plan means that the focal plane is a plane – the best for a camera sensor – instead of a sphere section), the magnification (for example  $60\times$ ) followed by the numerical aperture (for example 0.5), the tube length in mm ( $\infty$  means infinity corrected objective, other value is 160 for standard 160 mm tube length) followed by the thickness in mm of the glass slide placed between the lens and the sample for which the objective is corrected (usually 0.17 for the 0.17 mm thick coverslip used for biological sample, and a “-” when there is no such correction as is usually the case for MEMS observation) and sometimes at the bottom, the working distance in mm. Unfortunately different manufacturers use different ar-

rangement and sometimes, written figures become ambiguous (e.g., is 0.17 the NA or the glass slide thickness?).

The microscope subject is much richer than this short section exposes and for example we did not discuss the problem of illumination, including the possibility to use dark-field microscope for revealing small details on a flat surface. However, we shall finish the topic by noticing that we may beat the diffraction limit in an original way by using for imaging a very small aperture scanned very close to the sample instead of a lens. The short distance prevents diffraction effect to appear and the resolution becomes roughly equal to the aperture diameter - which can be down to 50 nm. With this so called Scanning Near-field Optical Microscope (SNOM) we need to scan line by line the object and record each points of the surface independently and sequentially, instead of the parallel way used for imaging with traditional light microscope. Even if it is slower and if its extremely short depth of focus is not easy to manage, the SNOM remains an optical microscope and as such renders important services for high resolution measurement of material optical properties.

### 3.8.2 SEM (Scanning Electron Microscope)

The scanning electron microscope (SEM) is the workhorse of 3D micro-characterization. Not only does it give hard to match resolution for geometry measurement (down to 1 nm), but also allows finding material composition or morphology. The SEM magic starts at the top of the vacuumed instrument column where electrons are produced (the so-called primary electrons – PE), accelerated and shaped in a beam before they are focused and scanned on the surface of the sample. Actually the SEM is part of the scanned microscopy (like the SNOM just described) where imaging is obtained point by point by progressively exploring the surface of the object using raster (succession of line) scanning. Here at the impinging point, the interaction between the primary electrons and the sample results in the generation of different signals that are measured with multiple detectors schematically shown in Figure 3.46.

The signals comes from different depth in the sample (from 10 nm to a few  $\mu\text{m}$ ), depending on the energy (or speed) of the primary electrons (from 200 eV to 30 keV) and on the type of interaction (Figure 3.47):

**Secondary Electron Imaging** uses the secondary electrons (SE) that are produced near the surface (<50 nm) of the sample where impinging electron are able to free electrons from the atoms shell, without them being recaptured. The number of electrons generated depends mostly on the topography of the image, with sharp slope being illuminated by a large number of electrons giving off more SE than flat surface. Actually the image retrieved in this mode are relatively simple to interpret (brighter is sloppier and darker is flatter) and because of the shallow interaction zone present the largest resolution that can exceed 1 nm.

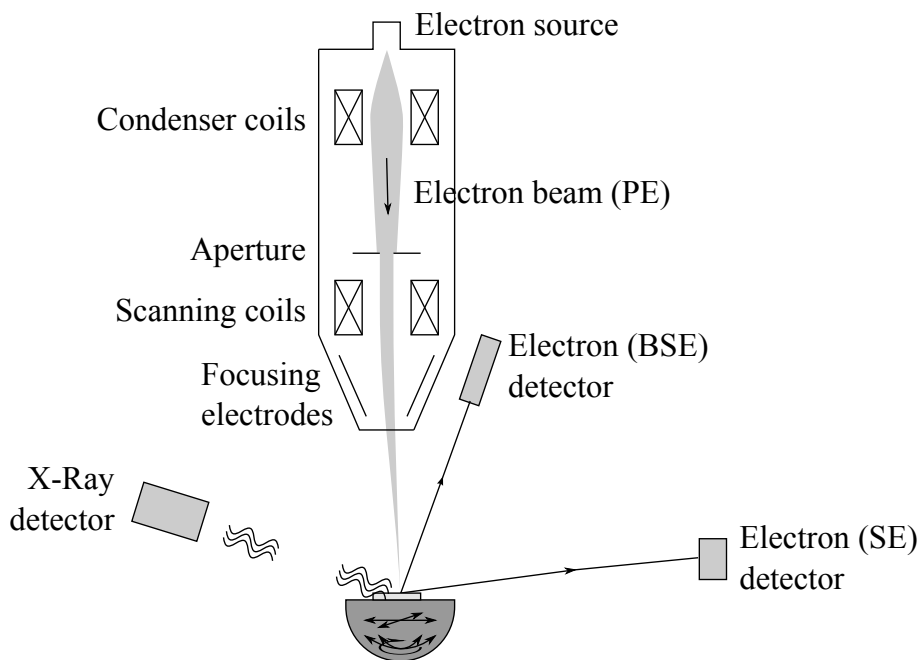


Figure 3.46: Simplified schematic of a typical scanning electron microscope.

**Back Scattered Electron Imaging** uses the back-scattered electrons (BSE) that are produced in a larger volume (1-2  $\mu\text{m}$ ) below the surface when the primary electrons recoil as they interact with the atoms. In this case the number of electrons returned depends directly on the atomic number of the element, the larger the Z number (heavier element) the brighter the image. This mode allows for comparative measurement of the composition of the sample where composition variation below 1% and down to a resolution of 10 nm can be resolved.

**X-Ray Spectroscopy** uses the X-rays that are produced when vacancy appearing in the inner shells of the atoms are filled. They originate from an even bigger volume (2-5  $\mu\text{m}$ ), thus with an even lower resolution, and the most common technique to study them is the energy dispersive X-ray spectroscopy (EDS). EDS allows for detecting qualitatively and sometimes quantitatively the atoms present in the sample for all elements between Boron and Uranium, the lighter elements presenting insufficient X-ray signature.

Actually, we also show in Figure 3.47 that if the sample is thin enough ( $< 100$  nm) and if we use higher energy beam (100 keV) transmitted electrons (TE) pass through the sample. By adding a detector below the sample we get measurement linked to absorption inside the material and we talk then of scanning transmission electron microscopy (STEM). However, the special configuration required for these measurement are best done inside a transmission electron microscope (TEM), which does not use scanning but projection and imaging for characteri-

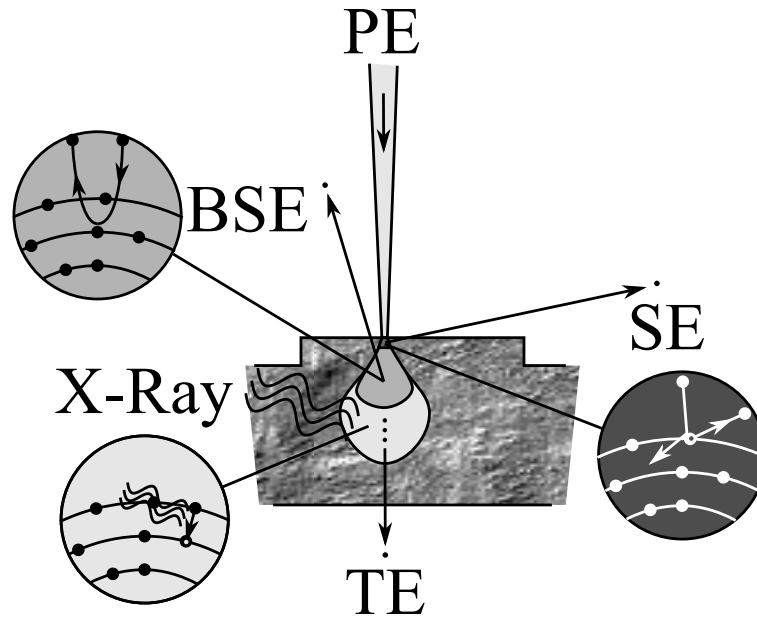


Figure 3.47: Interaction of electron with matter, showing primary electron beam (PE) with resulting backscattered electrons (BSE), secondary electrons (SE), transmitted electrons (TE) and X-rays.

zing the sample. In this case, by using ultra-thin samples (a few nm) it is even possible to reach an ultimate resolution well below the nm and ‘see’ individual atoms, revealing the atomic structure of the sample<sup>19</sup>.

The resolution of the SEM is directly linked to the size of the spot where the electrons can be focused and also on the interaction volume (cf. Figure 3.47). Contrary to the optical microscope, the diameter of the focused spot is not limited by diffraction effects. Actually the De Broglie’s wavelength of an electron at 30 keV is given by  $\lambda = h/p$  where the momentum  $p$  is actually obtained from the kinetic energy as  $p = \sqrt{2Em}$ , resulting in a value of...  $\lambda_e = 7$  pm, that is 0.007 nm. Even if the beam energy is 100 times smaller, the influence of diffraction (aperture would be 10’s of  $\mu\text{m}$  at least) can clearly be neglected. Thus in the SEM case, the spot size is still linked to the aberration of the electron optics (the coils, the electrostatic lens, etc), and is far from being “diffraction limited”. Currently the highest resolution claimed by manufacturer is below 1 nm and in some extreme case close to 1Å.

The main techniques for obtaining a better resolution would be to decrease the diameter of the aperture in the column (Figure 3.46) and to bring the sample closer to the column end. Both techniques reduces the number of stray electrons blurring the image and are also an effective means to lessen electron lens aberration. Of

<sup>19</sup>There are more signals coming from the electron excited sample, like the Auger electron studied with Auger electron spectroscopy (AES) or the photons from cathodaluminescence, that can be used for sample analysis, but fall beyond the scope of this short introduction.



course, in the first case, that means having a lower current reaching the sample and thus ‘darker’ images, but with modern detector it is usually not so much of an issue. These techniques aim at decreasing the SEM spot size, but it is also possible to decrease the interaction volume by lowering the velocity of the electrons and working with lower beam energy. Ultimately using in the focusing section of the column additional elements for performing ‘beam deceleration’ can decrease the energy of the electrons hitting the sample to about 50 eV, giving interaction depth as small as a few nm for secondary emission.

Another feature that differentiates the SEM is that it presents a very large depth of field, 10 to 100 times larger than what is typical with an optical microscope. Actually, for practical reason (focusing electrons is more complicated than photons), the NA of the focusing column is very small – for example, compare the length of a SEM column to the frontal distance of a microscope – and thus the illuminating beam has a very small divergence, keeping its diameter constant over a rather large distance, and hence the resolution of the SEM<sup>20</sup>.

An annoying feature of the conventional SEM, which operates in high vacuum, is that the specimen has to be electrically conductive or has to be coated with a thin conductive layer (e.g., carbon, Au), which is usually done inside a small sputter placed close to the tool. This avoids negative charge build-up on dielectric sample (electron get trapped in defect) that would finally prevent primary electrons to reach the surface, quickly decreasing image quality.

If reaching an ultimate resolution is not the main issue, the environmental scanning electron microscope (ESEM) circumvent this problem by proposing different vacuum modes besides the high vacuum. Actually instead of working in vacuum, the ESEM works with controllable pressure inside the chamber, while using a special pumping scheme for maintaining high enough vacuum in the SEM column. The presence of gas in the chamber absorbs the electron fast, thus the distance between the column and the sample has to be decreased and a new electron detector had to be designed for detecting the SE and BSE, but this does not affect significantly the resolution of the SEM. In the other hand, the gas atoms ionized in the chamber do a good job at neutralizing the electrons on the surface, completely removing the issue with charging effect on non-conductive samples that do not require metal deposition anymore. The possibility to have a high enough pressure in the chamber even allows for probing ‘wet’ samples, if not at room temperature at least by cooling it to decrease their water vapour pressure.

The large depth of field of the SEM is very interesting for imaging large MEMS with a relatively high resolution, which is unpractical with a standard microscope, and thus the main use of SEM will be for surface geometry observation and measurement. The observation is facilitated by using a sample holder with multiple rotation and translation axes (Figure 3.46), and then the depth may be estimated

---

<sup>20</sup>We may note that the term “depth of field” is inappropriate here as we don’t use the electron for imaging (except in the TEM), but only for very local “illumination” of the sample and then look in many direction what comes out of it.

by rotating the sample holder by 90° and looking to the sample sideways. However, accurate measurement can only be obtained by first sectioning the sample and looking perpendicular to the exposed surface, avoiding the projection angle distortion appearing with glancing observation.

Most SEM tools have a signal port allowing to power or control MEMS from outside the chamber allowing to test MEMS in operation. Still, applying electrical field inside the device will inevitably affect the electron path and may create artifact that may be hard to overcome, and when it is possible it is good to avoid large electrodes with high voltage and thus favor thermal actuator over electrostatic ones.

The EDS capability of the SEM is also very convenient to analyze locally the composition of materials and is useful for developing new material or checking the composition of unknown residue after processing.

### 3.8.3 Contact probe profilometry

The contact probe profilometry is a family of tools that are used for measuring surface profile by simply measuring the vertical displacement of a sharp tip that is scanned on the sample surface (Figure 3.48). The method is basically a line measurement method, but it can be used for surface measurement by repeating line scan in a raster like manner.

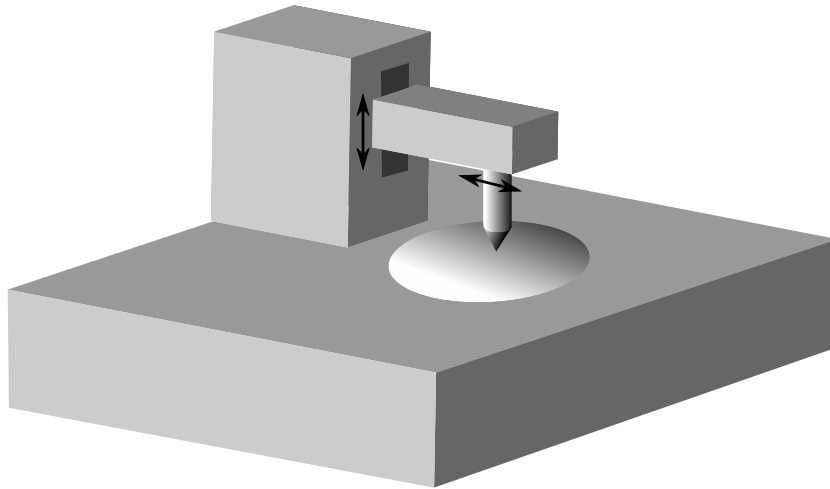
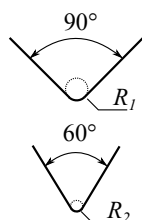


Figure 3.48: General principle of a contact-probe profilometer.

The stylus profilometer is the oldest of these tools, where a sharp metal tip is pressed against the surface and scanned while its vertical movement is recorded using magnetic sensing (LVDT). The user has only to set the sampling length (the distance of the measurement on the sample), the number of sampling points and the range of expected vertical motion, setting the vertical resolution of the measurement. In this way we easily get the profile of the surface, as shown in Figure 3.49, and by computing, for example, the height difference at points 1 and

3 we easily measure pit depth with high accuracy. However, it should be noted that important artifacts appear with the stylus profilometer when one tries to measure steep profiles: if the slope of the profile is steeper than the tip of the profilometer we can't record the surface profile.



Actually, a zoomed view of typical tips is shown in the inset, where we can see the rounded tip with varying radius of curvature (commonly 2  $\mu\text{m}$ , 5  $\mu\text{m}$  or 10  $\mu\text{m}$ ) and varying cone angles (normally 60° but 90° tip can also be found). We understand then why in Figure 3.49 at point 2 the slope profile is accurately measured, whereas at point 4 the slope is incorrectly recorded (actually we measure the tip slope) and at point 5 the narrowness of the hole makes the depth measurement wrong. To overcome this issue one may use sharper tip, but no tip is infinitely sharp and this limitation will always remain when high frequency corrugation appears on the sample surface. Another artifact that is of-

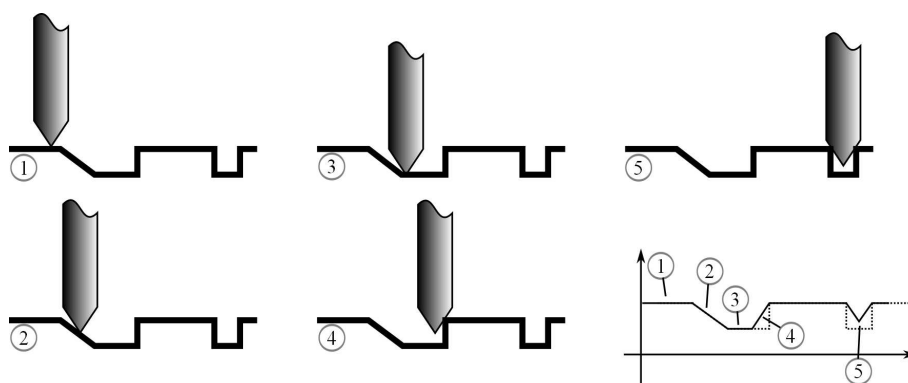


Figure 3.49: Measurement and artifacts with stylus profilometer.

ten observed with stylus profilometry is that measuring a bump or a hole profile from left-to-right or from right-to-left does not give the same profile. Actually the tip support has a tendency to bend when the tip goes up, changing the actual position of the tip and skewing the measured profile. For uncovering this effect we may scan the same feature from both directions and verify if the measured profile are the same. Finally, it should be noted that the tip is pressing over the sample surface with a force that is in the mN range, potentially leading to local deformation. This may affect substantially the recorded profile of soft materials like polymer.

A derivative of this tool is the atomic force microscope (AFM), where the tip is mounted on a very soft cantilever whose deformation is measured with an optical lever method. The high precision of the method allows measurement with high accuracy easily reaching sub 0.1 nm resolution while lowering the contact force, avoiding deformation or damage of sample surface. Moreover AFM are usually fitted with piezo-electric actuators for X-Y scanning, allowing accurate in-plane measurement with sub-nm resolution, compounded by the fact that the “bending

support problem” described above for stylus profilometer is controlled here. Of course, the actuators limit the achievable range of in-plane scanning that usually do not exceed 1 mm.

The contact type of measurement used by these tools implies that it is simple to interpret and it has become the reference tool for checking thin-film thickness after deposition. Of course this requires to create a step at a film edge for measuring and the trick here is to mask a part of the substrate with kapton tape or a drop of photoresist, make the deposition, and then lift-off the material atop the tape (lift the tape) or the photoresist (dissolve it in acetone), leaving a clear step at its edge.

The surface profile measured allows simple computation of many surface parameters like roughness, waviness or higher order parameters. Actually a surface profile can be considered as being a combination of undulations of different frequencies – it is after all a single variable signal, using spatial ( $x$ ) coordinate instead of the usual time variable  $t$ , where Fourier transform can be applied to get (spatial) frequency spectrum. The high frequency part is commonly called the roughness, while the mid-frequency range is the waviness. The measurement with a finite dimension tip appears as a low-pass filtering of the surface profile<sup>21</sup>, smoothing the recorded profile of the sample. A common parameter for evaluating roughness,  $Ra$ , is the arithmetic mean of the roughness profile

$$Ra = \frac{1}{l} \int_0^l |z(x) - \bar{z}| dx$$

where  $l$  is the sampling length and  $\bar{z} = \frac{1}{l} \int_0^l z(x) dx$  the average profile height. We need to understand that there is no unique value of this parameter, as a surface is mostly fractal, and using sharper and sharper tip will reveal more peak and valley. Thus the roughness is measured for a certain application dictating the useful range of spatial frequencies, and choosing the appropriate tool and data processing for achieving this goal. The high frequency cut-off (or low wavelength cut-off) is given by the tip radius and by the sampling period. Actually we can consider that the radius of curvature of the tip gives a “mechanical” cut-off wavelength of about the same value – as such a tip of 2  $\mu\text{m}$  radius will allow to measure down to a wavelength of 2  $\mu\text{m}$  (or a spatial frequency of 0.5  $\mu\text{m}^{-1}$ ). The sampling period – obtained by dividing the sampling length by the number of sampling points – gives rise to a cut-off wavelength which according to Nyquist theorem is at about twice its length. In practice, for recording roughness profile we would oversample the surface and use a sampling length about 1/6th of the desired cut-off wavelength. For example, for obtaining a cut-off only limited by the “mechanical” cut-off with a tip radius of 2  $\mu\text{m}$  thus resulting in a low wavelength cut-off of about 2  $\mu\text{m}$ , we should sample the surface every 0.35  $\mu\text{m}$ . At the other side of the spectrum, the high wavelength cut-off (or low frequency cut-off) will be given by the sampling length, and usually we would use 0.5 mm or so (i.e., spatial frequency of 2  $\text{mm}^{-1}$ ).

---

<sup>21</sup>More precisely it is a convolution of the sample and the tip profiles.

## Problems

- Plot using polar coordinate the Young's modulus variation w.r.t. to direction in the top surface plane for a  $\langle 110 \rangle$  Silicon wafer in a way similar to what was done for  $\langle 100 \rangle$  and  $\langle 111 \rangle$ -cut wafers in the inset p. 74.
- An optical telecommunication devices manufacturer wants to use microfabrication to produce V-groove for holding optical fiber. Which process can be used to achieve this goal ?
  - silicon substrate, photoresist mask and HF etchant
  - glass substrate, chromium mask and RIE etching with  $\text{SF}_6$
  - silicon substrate, silicon nitride ( $\text{Si}_3\text{N}_4$ ) mask and KOH etchant
  - silicon substrate, silicon dioxide ( $\text{SiO}_2$ ) mask and HF etchant

Draw a cut-out view of all the different processes proposed before making your choice

- A mask has the pattern of Figure 3.50(a). Which photoresist and pattern transfer technique could be used to pattern a thin film (in black) as shown in Figure 3.50(b)?

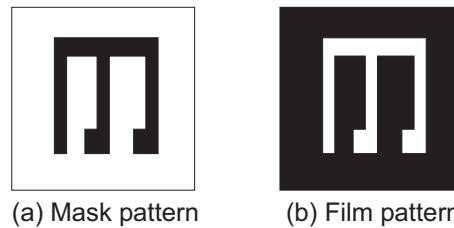


Figure 3.50: Mask and film pattern.

- positive photoresist and lift-off
  - negative photoresist and wet etching
  - positive photoresist and RIE etching
  - negative photoresist and lift-off
- A long slit in silicon is closed by growing oxide. The slit has a width of  $2\ \mu\text{m}$ .
    - Approximate the Deal and Grove's equation when  $t$  is much larger than  $\tau$  and  $A^2/4B$  (long duration approximation).
    - How long will it take to close the slit if the long duration oxidation is performed at  $1100^\circ\text{C}$  in wet  $\text{O}_2$ ? (Note: we have  $B = 0.51\ \mu\text{m}^2/\text{h}$  at  $1100^\circ\text{C}$ )

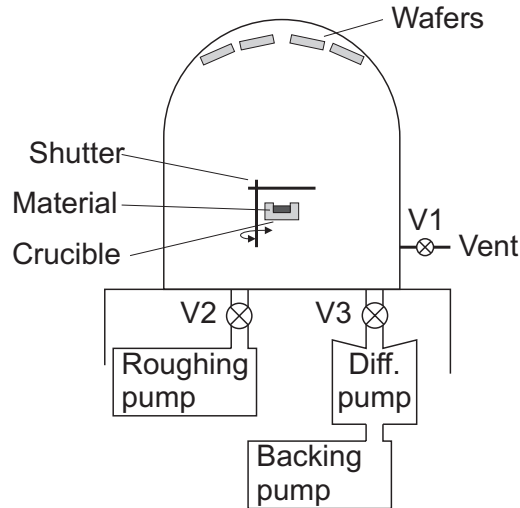


Figure 3.51: Evaporation system schematic.

5. An old evaporation system is schematically represented in Figure 3.51.
  - (a) The wafers are not placed in the same plane in the chamber. Justify their position using simple geometric argument.
  - (b) Suppose the chamber is originally under vacuum. Detail the sequence of operation needed to introduce the sample and prepare for evaporation (detail particularly the operation of the 3 valves).
  - (c) It is possible to make the vacuum system simpler by removing one of the pumps. Propose a new structure adding pipes and valves as needed and detail again the new sequence of operation to prepare for an evaporation.
  
6. Suggest a complete microfabrication process that can be used to fabricate the channel shown in Figure 3.52.

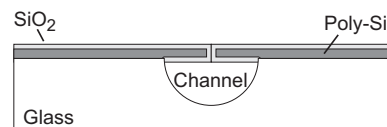


Figure 3.52: Sealed micro-channel on glass.

7. We consider the structure of Figure 3.53.
  - (a) Propose a complete process based on KOH etching that could be used to produce the structure.
  - (b) Justify the use of a sandwich of SiO<sub>2</sub> and Si<sub>3</sub>N<sub>4</sub>. Could the two layers order be inverted?

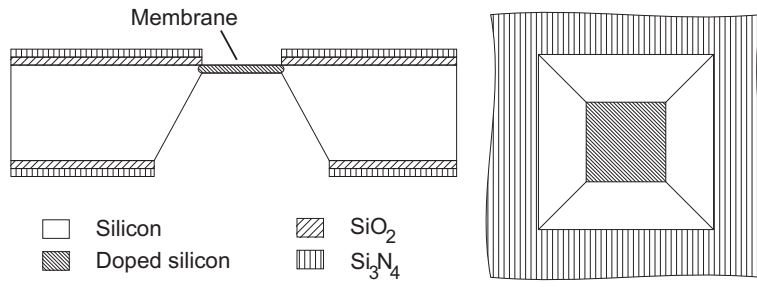


Figure 3.53: Membrane obtained by KOH etching.

8. Propose a fabrication process for producing the structure of Figure 3.54 . Indicate precisely the wafer type, the mask orientation and justify the slant in the beam.

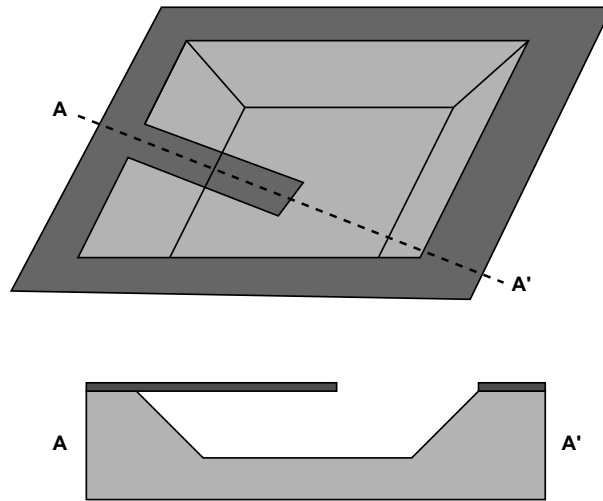


Figure 3.54: Slanted cantilever structure.

## Solutions

### Problem 4

1. The Deal and Groove's model gives

$$d_o = \frac{A}{2} \sqrt{1 + \frac{t + \tau}{A^2/4B}} - \frac{A}{2},$$

where  $B$  is the parabolic rate constant and  $B/A$  the linear rate constant. Typical values for these constant are  $\tau \approx 20$  min and for we oxidation at  $1000^\circ\text{C}$   $B = 0.35 \mu\text{m}/\text{h}$  and  $B/A = 1.8 \mu\text{m}^2/\text{h}$ , that is  $A \approx 0.19 \mu\text{m}^{-1}$ . Thus when  $t \gg \tau$  (that is after  $t > 10\tau \approx 3$  h) we have:

$$d_o \approx \frac{A}{2} \sqrt{1 + \frac{t}{A^2/4B}} - \frac{A}{2}$$

we also note that  $t \gg A^2/4B$  as  $A^2/4B = B/4(B/A)^2 \approx 0.027$  approx 1.6 min and we may ignore the 1 in the  $\sqrt{\quad}$  thus:

$$d_o \approx \sqrt{Bt} - \frac{A}{2}$$

We note even that in the condition considered  $t > 3$  h,  $\sqrt{Bt} > 1$  and  $\frac{A}{2} \approx 0.1 \mu\text{m}^{-1}$ , thus  $\sqrt{Bt} \gg \frac{A}{2}$ , and finally we may write that:

$$d_o \approx \sqrt{Bt}$$

This results shows why  $B$  is called the parabolic growth rate ( $d_o^2 \approx Bt$ ). This expression can be understood as a diffusion limited process, as oxygen has to diffuse through the oxide layer atop silicon. In fact instead of speaking of long time approximation, it would be more appropriate to talk of "thick layer" approximation: the simplified expression may be used as soon as  $d_o \geq 0.6 \mu\text{m}$ .

2. We will only consider in a first approximation the parabolic model and assume a simple 1D model. In that case to close the slit, we need to consider that at the same time the oxide is growing, silicon is consumed. In fact, when we have the growth of  $d_o \mu\text{m}$  of oxide we consume  $d_{\text{Si}} = 0.46d_o \mu\text{m}$  of silicon - it means that the interface advances by a total of :  $d = d_o - d_{\text{Si}} = 0.54d_o$ . That is, the oxide thickness needed for advancing by a distance  $d$  is  $d_o = d/0.54$ .

For closing the  $2 \mu\text{m}$  slit, the dioxide growth will happen from both sides of the slit, and thus we need to advance only by  $1 \mu\text{m}$  (we neglect here the extremities of the slit). Thus we need to grow:  $1/0.54 = 1.85 \mu\text{m}$  of oxide. Using the approximation derived previously ( $d_o < 0.6 \mu\text{m}$ ), we would need  $t = d_o^2/B = 6.7$  h, or about 6 hours and 43 minutes at  $1100^\circ\text{C}$  in wet  $\text{O}_2$ .



# Chapter 4

## MEMS technology

### 4.1 MEMS system partitioning

At the early stage of MEMS design an important question to be answered will be: hybrid or monolithic? Actually the decision to integrate the MEMS directly with its electronics or to build two separate chips has a tremendous impact on the complete design process. Most MEMS observer will advocate the use of separate chips and only in the case of a definite advantage (performance, size, cost) should a MEMS be integrated together with its electronics. From past industry examples,

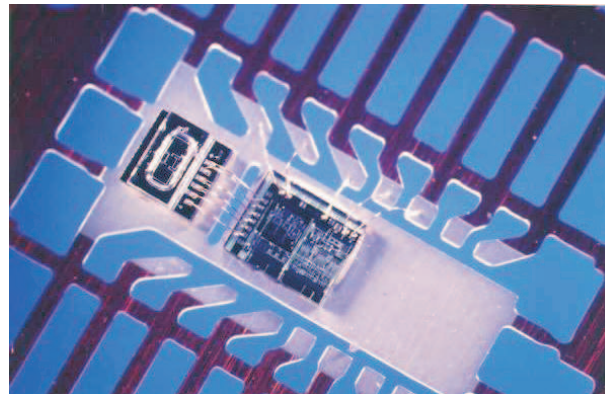


Figure 4.1: Hybrid integration in a pressure sensor (Courtesy Sensoror AS - An Infineon Technologies Company).

only a handful of companies, like Analog Device for its range of accelerometer or Motorola for its pressure sensors, have promoted the integrated process - and all are big companies having market reaching millions of chips. The hybrid approach in the other hand is used by many more companies on the market. For example Figure 4.1 shows a hybrid solution from Sensoror, the pressure sensor SP15. The MEMS chip on the left is connected to the ASIC on the right using tiny wire and both are mounted in a metal frame before encapsulation in the same package. The advantage of this solution, dubbed “system in the package (SIP)”, is that both

chips can use the best process without compromise and may achieve a better overall yield. However compactness and reliability suffers from the additional elements and the packaging becomes slightly more complicated. Moreover, the electronic is somewhat further from the sensing element and the tiny wires used for connection may introduce additional noise if the signal is small. It is this last argument that has pushed AD to develop its fully integrated accelerometer range, the iMEMS.

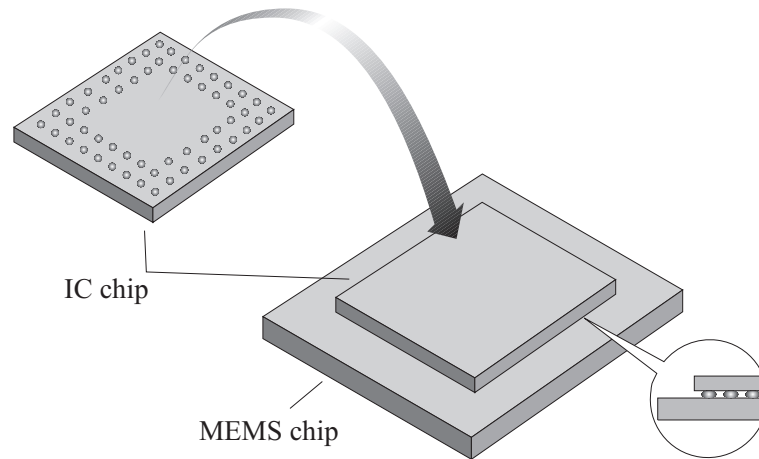


Figure 4.2: Flip-chip assembly of an IC chip on a MEMS chip.

More recently an intermediate solution has emerged that tries to reconcile both approaches: the flip-chip integration. In that case, shown in Figure 4.2, the electronics chip and the MEMS chip are interconnected using small solder balls that allows very tight assembly between them, with connection lead only a few  $\mu\text{m}$  long. This technique is very interesting, but it remains more expensive to use and has the drawback to hide the MEMS chip surface, preventing its inspection or open use.

By splitting the electronics design from the MEMS design we have made a step toward a simplification of the design, but more partitioning is required to help design complex MEMS structures. Actually, an interesting way to split the different sub-systems entering in the design of MEMS is to consider separately passive and active structures:

- Passive structures are used to support, guide, channel, etc providing indispensable building blocks for the realization of complete systems. Their main role is to transport energy within the system.
- Active structures are at the core of actuator and sensor operation. Fundamentally there role is to allow a certain form of energy transfer between the environment and the system.

But before we embark on a detailed review of the different technologies used in MEMS, we will have a view of an important limitation to our designs: the limitations imposed by the fabrication process capabilities, called the design rules.

## 4.2 Fabrication tolerance and design rules

Modeling and simulation does not give all the aspects of the MEMS device and at the layout stage the technology constraint needs to be taken into consideration. Actually there are many geometries that will give the same function (e.g. a spring constant of 1 N/m) but some are clearly better than others - in part because of fabrication tolerance - and some that simply won't be fabricated properly - and this could be prevented by following design rules.

### 4.2.1 Fabrication tolerance

In fact we have already discussed the impact of the fabrication process tolerance in the paragraph on relative manufacturing accuracy (Section 2.1). However we lacked experimental values to get a better feeling of what is possible for some typical MEMS processes.

The main tolerance we have to consider in micro/nanofabrication are the geometry tolerance, although we may sometimes be affected by material properties tolerance (residual stress, resistance, Young's modulus...).

The geometry tolerance gives the range of fabricated dimension (between different position on the same wafer, between different wafer in the same run, between different run) as a function of the layout dimension. It describes the fact that in a fabrication process the width of each patterned structure and the thickness of each layer have unavoidable random variations. The knowledge of this variation is a key element to perform design for manufacturability or '6s' process. Although, many fabrication processes are not fully characterized and the geometry tolerance is not known exactly, when it is provided, it has two parts:

**patterned width tolerance** that takes into account all the steps in the layer patterning process (mask fabrication, photolithography, pattern transfer). It is generally expressed as: Fabricated Dimension = Layout Dimension  $\pm_{\pm\min}^{\pm\max}$

**thickness tolerance** that takes into account the deposition uniformity over one wafer in the chamber and between different wafers from different runs.

Typical patterned width tolerance in MEMS process using contact mask photolithography will be  $+0.5 \mu\text{m}/-0.5 \mu\text{m}$  in your neighborhood clean-room - but it is seldom characterized and can be much worse. The Bosch MPW process, based on the growth of a thick epitaxial silicon layer as structural layer, gives for this layer the patterned width tolerance as Layout Dimension  $\pm_{-1.2 \mu\text{m}}^{-0.2 \mu\text{m}}$ . We note that this last tolerance is such that we always get narrower structures than expected, in the average by  $-0.7 \mu\text{m}$ . It is maybe what we need, but if this prove to be a nuisance, we may compensate automatically, without having to redraw the layout. In fact pattern generator may 'expand' automatically features just before producing the mask. By expanding the features by  $0.7 \mu\text{m}$ , we remove the systematic dimension

error and the tolerance will be  $\text{Layout Dimension}_{-0.5 \mu\text{m}}^{+0.5 \mu\text{m}}$ . Its amplitude did not change, but its mean value has been placed at 0.

If the possible patterned width variations induced by the tolerance pose a problem, we could use the fact that they often have absolute values: for a 2  $\mu\text{m}$  wide beam a tolerance of  $\pm 0.5 \mu\text{m}$  represent  $\pm 25\%$ , probably an issue if we want to know its spring constant, but the tolerance becomes only a mere  $\pm 5\%$  if the beam is 10  $\mu\text{m}$  wide.

For thickness tolerance, depending on the technique used, it may vary in your neighborhood academic cleanroom between  $\pm 2\%$  for the best techniques like LP-CVD of PolySi (even down to  $\pm 1\%$  for evaporation of Al with mask compensation in a large chamber), to close to  $\pm 10\%$  for some sputtering process. In general CVD techniques tends to achieve better results than PVD techniques in this aspect. The Bosch MPW process gives a thickness tolerance of  $10.6 \mu\text{m} \pm 1.5 \mu\text{m}$  for the epitaxial layer at the core of the structure. The MUMPS process is a surface machining process with 3 structural polysilicon layers, 2 silicon dioxide sacrificial layer and a final metal layer. MEMSCAP, who runs the process, gives the thickness tolerance of the different layers deposited as shown in Table 4.2.1. As we see here the

Film	Thickness [nm]			Residual stress [MPa]			Resistance [ $\Omega/\text{sq}$ ]		
	min.	typ.	max.	min.	typ.	max.	min.	typ.	max.
Nitride	560	600	640	0	90	180	-	-	-
Poly0	460	500	540	0	-25	-50	15	30	45
Oxide1	1900	2000	2100	-	-	-	-	-	-
Poly1	1950	2000	2050	0	-10	-20	1	10	20
Oxide2	710	750	790	-	-	-	-	-	-
Poly2	1425	1500	1575	0	-10	-20	10	20	30
Metal	4800	5200	5600	0	50	100	0.05	0.06	0.07

Table 4.1: Typical thickness, residual stress and resistance tolerance in the layers of the MUMPS process.

thickness tolerance varies between  $\pm 8\%$  and  $\pm 2.5\%$ , with CVD (Nitride, PolySi, Oxide) performing generally better than PVD (Metal), except for the thin Poly0 layer. The table also add the tolerance on other material properties, that will give a better view of what can be expected after a process is run.

## 4.2.2 Design rules

To help the designer in its design and at least get functional devices - even if they don't work exactly as intended because of fabrication tolerance - the process comes

---

**Example 4.1** Effect of tolerance in the Bosch MPW process.

---

WE ARE CONSIDERING a  $200\ \mu\text{m} \times 200\ \mu\text{m}$  square membrane for a pressure sensor. We will use the Bosch MPW process for fabrication and use its thick epitaxial silicon layer for the membrane. For this process the patterned width tolerance is given by  $\text{Layout Dimension}_{-1.2\ \mu\text{m}}^{-0.2\ \mu\text{m}}$  and the epitaxial layer thickness tolerance is given as  $10.6\ \mu\text{m} \pm 1.5\ \mu\text{m}$ .

The deflection of a square membrane under uniform pressure is given by  $y_C = \frac{\alpha P a^4}{E t^3} p$ , where  $a$  is the side of the membrane,  $t$  its thickness,  $E$  the Young's modulus of the silicon and  $p$  the pressure. It means that the absolute value of the deflection relative error is given by  $\frac{\Delta y_C}{y_C} = 4 \frac{\Delta a}{a} + \frac{\Delta E}{E} + 3 \frac{\Delta t}{t} + \frac{\Delta p}{p}$ . Focusing on the error due to geometry tolerance (the Young's modulus tolerance is very low as it is obtained by epitaxy), we have:

$$\frac{\Delta y_C}{y_C} \approx 4 \frac{\Delta a}{a} + 3 \frac{\Delta t}{t}$$

For the known process tolerance ( $\frac{\Delta a}{a} \approx 0.5\%$  and  $\frac{\Delta t}{t} \approx 15\%$ ) it means that the relative tolerance on the deflection reaches  $47\%$ . Said another way, the deflection could in practice be as small as  $32\ \text{nm}$  or as large as  $100\ \text{nm}$ , for the same layout of the mask, the same Bosch Foundry process... just because of process variability! This most probably implies that our sensor will have to use calibration if it wants to be of any use.

---

usually with design rules. Following these rules does not guarantee the device *will* work (there are too many ways to make quirky things...) but it will strongly increase the odds it does. In fact, failing to follow the design rules will generally results in device that works on paper... only!

The rules can be split in three categories:

- minimum dimension rules for each mask level
- registration rules between two mask levels
- rules for other limitations of the process

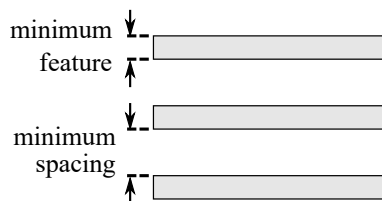


Figure 4.3: Minimum dimension rules.

The minimum dimension rule (Figure 4.3) is linked mostly to patterning process resolution and it comes in two flavors:

**Minimum feature rule** the smallest width that a structure can have to be certain it is actually fabricated – thinner structure in the layout may not appear after fabrication.

**Minimum spacing rule** the smallest gap that could exist between two structures in the same layer to be sure they won't be touching each other

For typical MEMS process in academic cleanroom the minimum feature/spacing rule is often 2  $\mu\text{m}$  (or more rarely 1  $\mu\text{m}$ ) – much larger than in IC fabrication. In fact the thickness of the layer used in MEMS and the common use of contact mask aligner prevent patterning structure much smaller than 1  $\mu\text{m}$ . Of course using nanolithography tools (e-beam exposure, FIB patterning...) changes this point and could allow to reach dimensions down to a few nm - but don't try it on a full wafer.

For design with multiple mask levels the set of registration rules concern the alignment between the patterns on the different pair of levels of masks. These rules

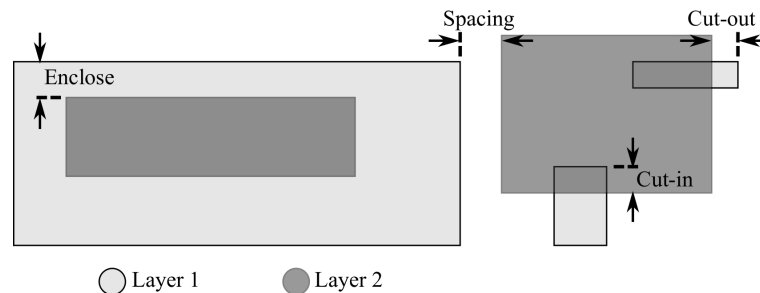


Figure 4.4: Registration rules.

helps to be sure that patterns on two different levels of mask have the expected position one with respect to the other. They are described as:

**Enclose** this rule makes sure the pattern of layer 1 encloses the pattern in layer 2

**Spacing** this rule makes sure the pattern of layer 1 does not contact the pattern in layer 2

**Cut-in** this rule makes sure the pattern of layer 1 finishes under the pattern in layer 2

**Cut-out** this rule makes sure the pattern of layer 1 finishes outside the edge of the pattern in layer 2

We give in Table 4.2.2 the registration rules in the MUMPS process for the Poly2 mask level. As can be seen there, many numbers are unspecified because they are very unlikely to be of interest. It should be noted that the MUMPS design handbook also gives rules for easy design, ensuring the device will be fabricated as

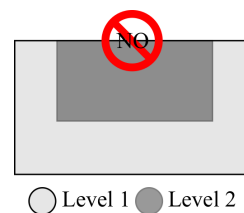
Level 1	Level 2	Enclose	Spacing	Cut-in	Cut-out
	Poly0	-	-	-	-
	Poly1	-	3 $\mu\text{m}$	5 $\mu\text{m}$	4 $\mu\text{m}$
Poly2	Via	4 $\mu\text{m}$	-	-	-
	Anchor2	5 $\mu\text{m}$	-	-	-
	Metal	3 $\mu\text{m}$	-	-	-

Table 4.2: Registration rules for Poly2 masks with respect to the other mask level in the MUMPS process (‘-’ means unspecified).

intended without too much fuss, and in that case it suggests to use all rules with a value of 5  $\mu\text{m}$ .

In the registration rules illustration in Figure 4.4, one may be surprised to see that there is no example of superimposed *parallel* pattern edge on the two levels as shown in the inset here.

In fact, the registration tolerance (which is not much better than 1 or 2  $\mu\text{m}$  in academic cleanrooms) makes such ‘perfectly’ aligned patterns impossible to fabricate. Accordingly the designer has to choose (the edge of pattern in level 1 *has to be* on which side of the edge of pattern in level 2?) and use the appropriate design rules (Enclose, Spacing, Cut-in or Cut-out) to be sure the fabricated device is as intended.



For design, we need not only to consider the rules that gives limitation on the dimension and on registration but we also need to respect extra rules to get working devices. We have already given some details on process limitations for a multi-layer surface micromachining (Section 3.5.2), which will need to:

- use release holes on wide plate structures to ensure sacrificial layers will be easily removed (MUMPS process impose a 30  $\mu\text{m}$  maximum Spacing value between holes),
- take measure to prevent stiction if the release step is to be conducted in a wet environment,
- take care of layer interaction between stacked layers,
- don’t forget to place a grounding plane below the mobile structure to avoid charge induced stiction during operation (and use bipolar excitation),
- make a T pattern and not a L pattern when two beams connect to avoid rounding of the external corner,
- ...

Only experience (its own or from others) will tell the details of what needs to be done here, and unfortunately, in academic clean-rooms, the difficulty to document the design pitfalls pushes the burden of ‘making it work’ on the micro/nanofabrication process that too often becomes a trial and error process...

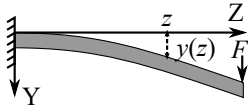
But, be sure that if the design rules are well documented - the design rules !

## 4.3 Passive structures

### 4.3.1 Mechanical structures

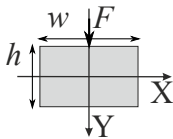
Most MEMS required the development of special micro-mechanical elements to achieve standard mechanical functionalities as linkage, spring, articulation, etc.

The spring, because traditional microfabrication of coiled springs is generally too complex to be considered, has seen many original developments based on the elastic properties of beams.



Actually the elasticity of a simple beam is well known [9], and for a beam submitted to a pure internal<sup>1</sup> transverse bending moment  $M$  the deflection  $y(z)$  is governed by:

$$\frac{d^2y}{dz^2} = -\frac{M}{EI} \quad (4.1)$$



with  $E$  the Young's modulus for the material of the beam and  $I$  the second moment of inertia for the beam cross-section, which is given by  $I = \int x^2 dA = wh^3/12$  for a beam with a rectangular cross-section as shown in the inset.

For the cantilever of the inset submitted to a point load normal to the surface at its end, the moment is simply  $M(z) = F(L - z)$  and the beam equation becomes :

$$\frac{d^2y}{dz^2} = -\frac{F}{EI}(L - z)$$

we integrate twice w.r.t.  $z$  the equation and obtain:

$$y = -\frac{F}{EI} \left( \frac{1}{2}Lz^2 - \frac{1}{6}z^3 + Az + B \right)$$

Considering the boundary conditions at  $z = 0$  :

- deflection is null  $y(0) = 0$ , thus  $B = 0$
- slope is null  $\left. \frac{dy}{dz} \right|_{z=0} = 0$ , thus  $A = 0$

giving finally

$$y = -\frac{F}{EI} \left( \frac{L}{2}z^2 - \frac{1}{6}z^3 \right) = -\frac{Fz^2}{6EI}(3L - z)$$

<sup>1</sup>The internal moment is exactly balanced by an externally applied moment of opposite sign.



Type	Deflection	Max Defl.	Spring constant
Cantilever	$y = \frac{Fz^2}{6EI}(3L - z)$	$y(L) = \frac{FL^3}{3EI}$	$\frac{3EI}{L^3}$
Clamped-guided beam	$y = \frac{F}{12EI}(3Lz^2 - 2z^3)$	$y(L) = \frac{FL^3}{12EI}$	$\frac{12EI}{L^3}$
Clamped-clamped beam	$y = \frac{F}{192EI}(12Lz^2 - 16z^3)$	$y(L/2) = \frac{FL^3}{192EI}$	$\frac{192EI}{L^3}$

Table 4.3: Characteristics of beams in bending.

The deflection at the end of the beam where the force is applied is thus

$$y(L) = -\frac{FL^3}{3EI}$$

giving an equivalent spring constant :

$$k = \frac{F}{y(L)} = \frac{3EI}{L^3}$$

We list in Table 4.3 the deflection when a point force is applied on the beam for

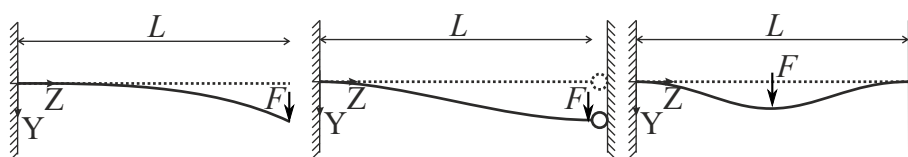


Figure 4.5: Beams with different boundary conditions (cantilever, clamped-guided beam, clamped-clamped beam) in bending.

different cases detailed in Figure 4.5. We also add the expression of an equivalent spring constant, corresponding to the deflection at the point of application of the force divided by the force. This parameter is useful for simple 1D lumped parameter modeling, where the beam is considered to be an ideal unidirectional spring.

But of course these beams are not enough and need usually to be combined together for forming suspension with added flexibility. Actually for choosing a suspension there are usually four main characteristics to watch: the spring constant in the direction of use, the compliance in the other directions (it needs to be low to keep the motion in the desired direction), the tolerance toward internal stress (long beam may buckle during fabrication) and the linearity during large deformation. There is of course a trade-off to be observed and different designs have been pursued (Figure 4.6) which have the characteristics shown in Table 4.4. As we see here, the folded beam suspension is versatile and particularly suitable for process where there is a risk of buckling (it will stand large internal stress, as those appearing in surface micromachining) but other designs may be suitable with other processes.

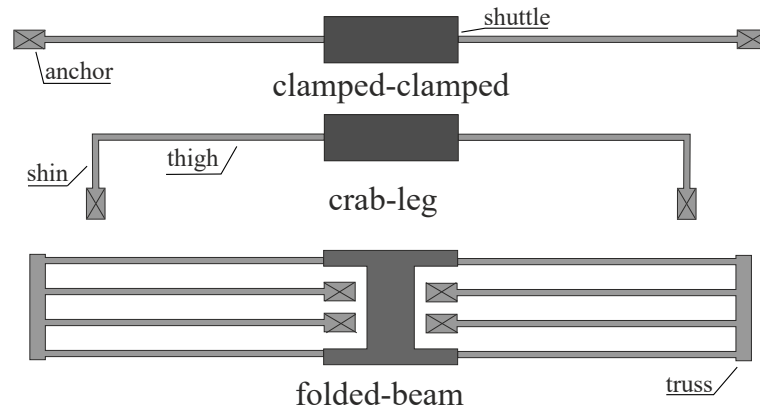


Figure 4.6: Layout of different type of suspension.

Type	Compliance	Buckling	Linearity
clamped-clamped	++	-	-
crab-leg	controllable	0	0
folded-beam	+	+	+

Table 4.4: Comparison of typical MEMS suspensions.

For example a clamped-clamped beam could be preferred in process with little or no internal stress like DRIE micromachining where structure is etched in bulk Si wafer, as it would allow building more compact suspension.

In the case of suspension made of multiple beams the computation of the spring constant may be complicated but often existing symmetries allow to decompose the suspension into elementary beams connected in series and in parallel. For two beam connected in series, the equivalent spring constant is simply the sum of the two beams spring constants whereas if the two beams are in parallel, the resulting spring constant is the inverse of the sum of the inverse of the spring constants - in fact the spring constant behaves in a similar way as capacitor in electronic circuit. For example, a clamped-clamped beam can be seen as two clamped-guided beams of half length in series, while a clamped-guided beam can also be decomposed in two cantilevers of half length connected in series. Accordingly, the spring constant of the clamped-clamped beam and the clamped-guided beam are, respectively,  $1/(4 \times 1/(1/4)^3) = 64$  times (i.e., 4 cantilevers of length  $L/4$  in series) and  $1/(2 \times 1/(1/2)^3) = 4$  times (i.e., 2 cantilevers of length  $L/2$  in series) the cantilever spring constant, as indicated in Table 4.3.

Of course the force on the beam is not always concentrated and in some application, particularly in fluidic application, the beam is subject to a uniform line pressure,  $q$ , in N/m. In that case it is less relevant to define an 'equivalent spring constant' (we note, still, the force on the beam is  $qL$ ) but the deflection can still be

obtained and is given in Table 4.5 corresponding to the cases shown in Figure 4.7.

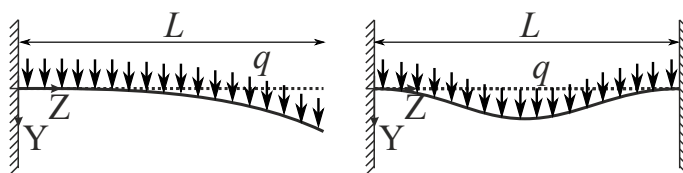


Figure 4.7: Beams under pressure.

Type	Deflection	Max deflection
Cantilever	$y = \frac{qz^2}{24EI}(z^2 - 4Lz + 6L^2)$	$y(L) = \frac{qL^4}{8EI}$
Clamped-clamped beam	$y = \frac{qz^2}{24EI}(z^2 - 2Lz + L^2)$	$y(L/2) = \frac{qL^4}{384EI}$

Table 4.5: Deflection of beams under pressure.

In addition to beams, MEMS often uses membranes or diaphragms, that is plates with thickness at most 20% of their width, for which the behaviour equations are usually more complicated. We give here the simpler case of small deflection for round and square membranes clamped at the edge (Figure 4.8)<sup>2</sup>. These cases are corresponding to typical application for pressure sensor or valves, showing the general dependence of the characteristics with the geometry (Table 4.6).

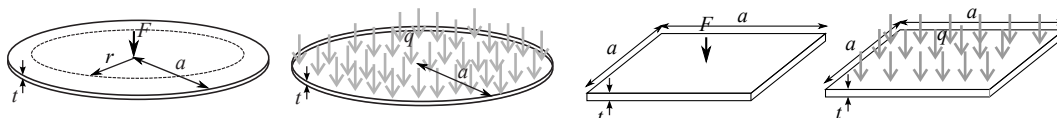


Figure 4.8: Deflection of round and square membranes under point force and uniform pressure.

Of course the case of deflection of membrane under pressure is of particular interest for pressure sensors, where the measurement of the deflection or the stress will be used to infer the pressure. We note that the formulas given are for small deflection, and are not valid for large deflection where bending stress can not be neglected. For example for a round diaphragm under pressure  $q$  with large deflection (assuming the material remains in its elastic limit) we obtain the following

<sup>2</sup>In order to remain linear, diaphragms are limited to deflection of about 30% of their thickness  $t$  and in that case experience only tensile stress. Membranes allows large deflection -  $y_C/t > 5$  - but are made of a soft material (low flexural rigidity), showing no bending stress and only tensile stress.

Type	Deflection	Max Defl.	Spring constant
Round (Force)		$y_C = \frac{Fa^2}{16\pi D}$	$k = \frac{16\pi D}{a^2}$
Round (Pressure)	$y = \frac{q}{64D}(a^2 - r^2)^2$	$y_C = \frac{qa^4}{64D}$	
Square (Force)		$y_C = \frac{\alpha_F Fa^2}{Et^3}$	$k = \frac{Et^3}{\alpha_F a^2}$
Square (Pressure)		$y_C = \frac{\alpha_P qa^4}{Et^3}$	

Table 4.6: Deflection of round and square membrane (Plate constant  $D = Et^3/[12(1 - \nu^2)]$ ,  $\alpha_F = 0.014$  ( $\nu = 0.3$ ),  $\alpha_P = 0.061$  ( $\nu = 0.3$ ).)

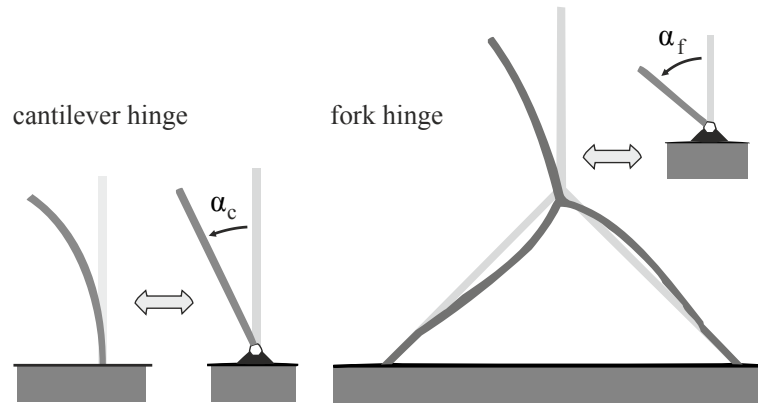


Figure 4.9: Flexure hinge and equivalent free hinge for (left) standard cantilever hinge (right) improved fork hinge

non-linear characteristic equation to obtain the center deflection  $y_C$ :

$$q = 64 \frac{D}{a^4} y_C + 4(7 - \nu) \frac{D}{t^2 a^4} y_C^3$$

Besides suspension, other mechanical function, like hinge or joint, are often needed in MEMS. However the fundamental inability to miniaturize hinge because of the low relative manufacturing accuracy evoked previously, forces designers to often use flexible micro-joint instead. These joints will have excellent wear characteristic but they will usually restrict rotation. These flexure hinge may use a simple cantilever or more complex beam arrangement. As we can see schematically in (Figure 4.9), the fork hinge[27] has the advantage to present a larger rotation angle for the same horizontal displacement than a standard cantilever beam with the same stability (resistance to buckling). Of course, if the angle of rotation need to be really large ( $> \pm 20^\circ$ ), the alternative is to use a free hinge as shown in Figure 3.33, but the manufacturing complexity will increase substantially - and the reliability will drop.

### 4.3.2 Distributed mechanical structures

Using lumped elements to represent continuous structure is clearly an approximation, and it requires good judgment to decide how a structure should be represented. For example, as we see in Figure 4.10, a beam can be represented by a single lumped element, a spring, or we can also take into account the weight of the beam itself and represent it by a lumped spring and a lumped mass. We could even go further, and consider the material loss inside the beam (linked to atomic rearrangement during deformation) and add to the model a dash pot to represent it as a viscous loss. The choice between the three representation illustrated in

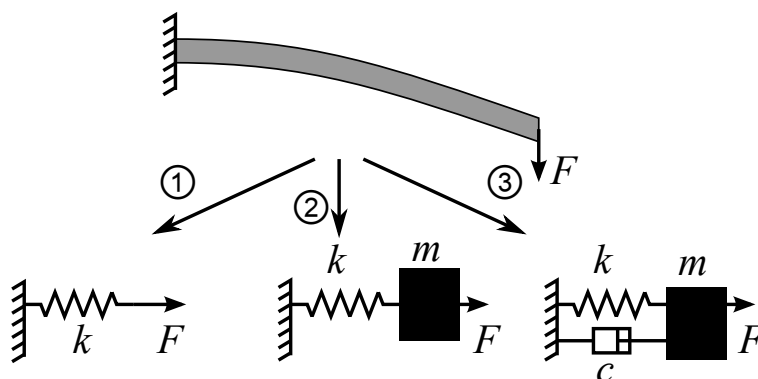


Figure 4.10: Modeling a continuous beam using different approximation with lumped elements.

Figure 4.10 will directly depend on an estimation of the dominant effect in the complete structure. For example if a mass is connected at the tip of the bending beam, and the mass is 20 times the weight of the beam itself, the inertia of the beam can probably be ignored, and its mass forgotten in the model. Likewise, if we place the beam in vacuum and damping due to air becomes negligibly small, then material damping could become important and a dash pot could be added to the beam model.

In the case of an elastic structure with a distributed mass, a method pioneered by Lord Rayleigh can be used to estimate an equivalent mass for the moving structure while keeping the spring constant obtained from beam or shell theory. The Rayleigh method is based on the hypothesis that at resonance (where there are no loss in the system) the maximum kinetic energy and the maximum potential elastic energy are equal.

The method is best explained using an example and we will consider a shuttle mass with a clamped-clamped suspension, composed, by symmetry, of two clamped-guided beams, as shown in Figure 4.11.

We need first to estimate the kinetic energy in the structure when it is excited sinusoidally at resonance with  $\omega = \omega_0$ .

We consider one beam of the suspension and approximate the amplitude of its resonant mode by using the expression of the static deflection with a force at the

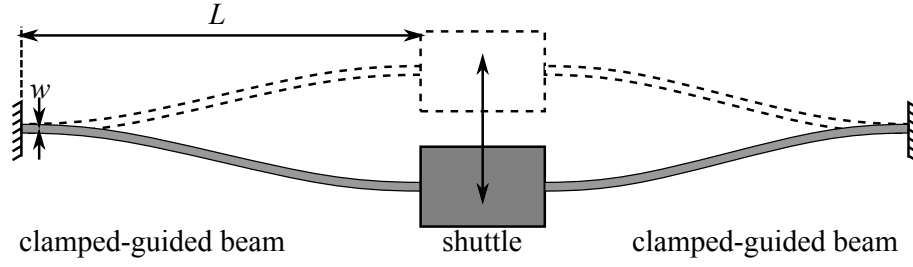


Figure 4.11: A shuttle mass oscillating between two positions with a pair of clamped-guided beams of length  $L$ , width  $w$  and thickness  $h$  as suspension.

end as given in Table 4.3.

$$y(z, t) = \frac{F}{12EI}(3Lz^2 - 2z^3) \sin(\omega t)$$

We take the derivative to obtain the velocity distribution along the beam as:

$$v(z, t) = \dot{y}(z, t) = \frac{F}{12EI}(3Lz^2 - 2z^3)\omega \cos(\omega t)$$

The kinetic energy ( $\frac{1}{2}mv^2$ ) in the beam is then obtained by integrating over the volume:

$$K_b = \int_0^L \rho h w v^2(z, t) dz = \frac{13}{70} \rho h w L^7 \left( \frac{F}{12EI} \omega \cos(\omega t) \right)^2$$

We will now turn our attention to the determination of the potential elastic energy inside the beam.

The elastic energy density in pure bending (that is, we have  $\sigma = -My/I$ ) for isotropic materials is given by:

$$\frac{1}{2} \sigma \epsilon = \frac{1}{2} \frac{\sigma^2}{E} = \frac{1}{2} \frac{(-My/I)^2}{E}$$

Noting that the bending moment along the beam is given by

$$M(z, t) = EI \frac{d^2 y(z, t)}{dz^2} = \frac{-F}{12} (6L - 12z) \sin(\omega t),$$

we can obtain the potential energy in the complete beam by integrating the energy

density over the beam volume:

$$\begin{aligned}
 U_b &= \int_V \frac{1}{2} \sigma \epsilon dV = \int_V \frac{1}{2} \frac{(-My/I)^2}{E} dV \\
 &= \frac{1}{2EI^2} \int_V M^2 y^2 dV = \frac{1}{2EI^2} \int_0^L M^2 dz \int_A y^2 dA \\
 &= \frac{1}{2EI^2} \int_0^L M^2 dz I = \frac{1}{2EI} \int_0^L M^2 dz \\
 &= \frac{1}{2EI} \int_0^L \left( \frac{-F}{12} (6L - 12z) \sin(\omega t) \right)^2 dz \\
 &= \frac{F^2 L^3}{24EI} \sin^2(\omega t)
 \end{aligned}$$

At resonance  $\omega = \omega_0$  we have a periodic transfer between kinetic and potential energy (for non dissipative systems) thus the maximum of kinetic energy is equal to the maximum of potential energy:

$$\begin{aligned}
 \max K_b &= \max U_b \\
 \Rightarrow \frac{13}{70} \rho h w L^7 \left( \frac{F}{12EI} \omega_0 \right)^2 &= \frac{F^2 L^3}{24EI}
 \end{aligned}$$

and we get

$$\omega_0^2 = \frac{70}{26} \frac{12EI}{\rho h w L^4}$$

If the continuous beam can be represented by a massless spring and a punctual mass at its end (lumped model), we have:

$$\omega_0^2 = \frac{k_{\text{eff}}}{m_{\text{eff}}}$$

where  $k_{\text{eff}}$  the effective spring constant is taken as the spring constant of a single clamped-guided beam as defined by beam theory and  $m_{\text{eff}}$  is obtained by identification with the previous formula.

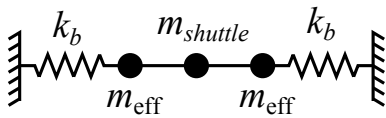
Thus we have for the equivalent model of the beam a spring of stiffness

$$k_{\text{eff}} = k_b = \frac{12EI}{L^3}$$

and a load of mass

$$m_{\text{eff}} = \frac{k_{\text{eff}}}{\omega_0^2} = \frac{26}{70} \rho h w L = \frac{26}{70} m_b$$

where  $m_b$  is the mass of the beam.



Thus the complete suspension we are considering composed of two equal springs and a central mass can be represented by this equivalent model.

It means, for example, that its resonant frequency could be estimated using  $\omega_0 = \sqrt{k/m}$  with  $k = 2k_b$  and  $m = m_{shuttle} + 2\frac{26}{70}m_b$ .

### 4.3.3 Fluidic structures

In microfluidic devices the ubiquitous passive element is the channel which is used to transport fluids. But before treating the fluid dynamics we need to have a look at the static properties of fluids at microscale.

From a physical point of view the forces between the solid and the fluid and inside the fluid are linked to attractive van der Waals forces. These forces arise at the molecular level because of electrostatic interaction and, particularly for neutral molecules, between induced dipoles at very short distance. Additionally one should take into consideration the Pauli exclusion principle that would prevent electronic orbital to interpenetrate and give rise to a strong repulsive force when the molecules becomes too close. These forces are semi-empirically described as

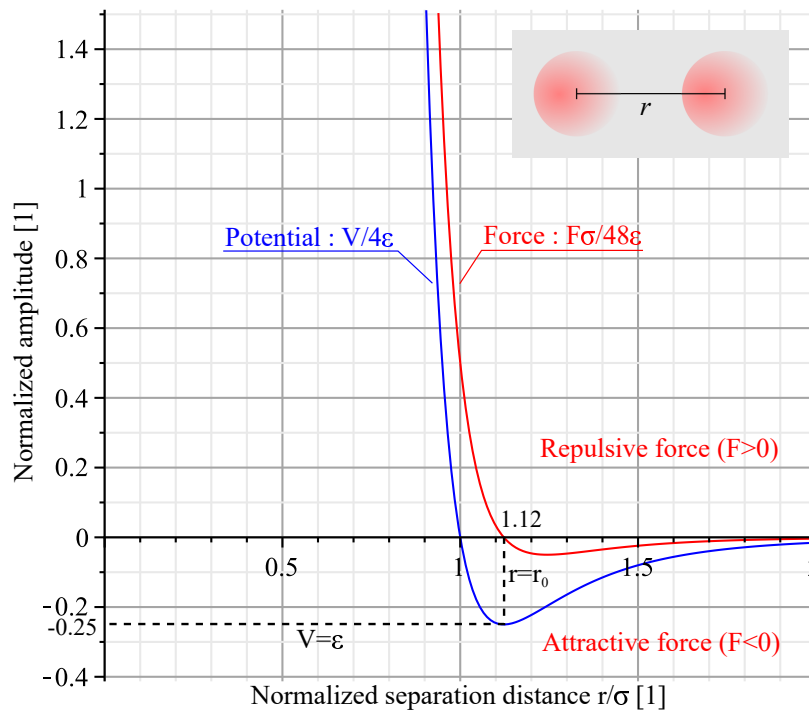


Figure 4.12: Force and potential energy between non-polar molecules as a function of their separation distance.

deriving from the Lennard-Jones potential

$$V(r) = 4\epsilon \left[ \left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^6 \right]$$



with  $r$  the separation distance between the molecules,  $\epsilon$  giving the potential energy well depth and  $\sigma$  the separation distance where the potential energy is 0. The power of 6 dependence of the attractive component is obtained from a physical model, of the different components of the attractive van der Waals forces between polar and non-polar molecules : the London dispersive force between two non-polar molecules, the Debye force between one polar and one non-polar molecules and the Keesom force between polar molecules. In practice, the London dispersive component is dominating except for strongly polar liquid like water, and its order of magnitude is  $6-7k_B T$  when the molecules are in ‘contact’ (that is, at the distance corresponding to the minimum of Lennard-Jones potential energy). In the other hand, the power of 12 of the repulsive part modeling the Pauli exclusion is somewhat arbitrary and used to decrease the computing complexity – as we already have the power of 6 we just need to square this result. For this part it has been proposed that an exponentially varying potential would be more appropriate, but it does not change substantially the results discussed here and shown in Figure 4.12. Classically, the corresponding force is derived as :

$$\vec{F}(r) = -\frac{\partial V}{\partial r} \hat{r} = 48 \frac{\epsilon}{\sigma} \left[ \left(\frac{\sigma}{r}\right)^{13} - \left(\frac{\sigma}{r}\right)^7 \right] \hat{r}$$

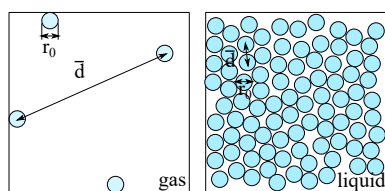
where  $\hat{r}$  is a unit vector on the  $r$  axis.

---

**Example 4.2** Molecules in liquid and gas nitrogen.

---

THE INTERMOLECULAR distance in (perfect) gas like  $N_2$  may be obtained from the molecular density. At standard pressure (101325 Pa) and temperature ( $0^\circ\text{C}$ ) the volume of one mole of molecule (molar volume) is given by  $V_M = RT/P = 22.4$  liter. That gives a density of  $N_A/V_M$  molecules/ $\text{m}^3$ , or differently said, one molecule occupies a (cubic) volume of  $V_M/N_A$  whose side length roughly gives us the intermolecular distance:  $\bar{d} = \sqrt[3]{V_M/N_A} = \sqrt[3]{RT/PN_A}$ . We can simplify this expression by recalling that  $R = k_B N_A$  giving  $\bar{d} = \sqrt[3]{k_B T/P}$ . Thus  $\bar{d} = 1.38 \cdot 10^{-23} \cdot 273.15/101325 = 3.33$  nm for a perfect gas at standard pressure and temperature. In the other hand the parameters for  $N_2$  in the Lennard-Jones based model are  $\epsilon/k_B = 91$  K and  $\sigma = 0.368$  nm, giving  $r_0 = \sqrt[6]{2}\sigma = 0.41$  nm – the molecules are roughly in the average 8 ‘diameters’ apart.



In the case of liquid nitrogen at 77 K we find a density of  $\rho = 800$   $\text{kg}/\text{m}^3$  and molar mass of  $m_M = 28$  g/mol (nitrogen molecule retains its diatomic nature upon liquefaction). It means that we have  $\rho/m_M$  mol/ $\text{m}^3$  or  $\rho N_A/m_M$  molecules/ $\text{m}^3$ . Differently said, one molecule occupies a volume of  $m_M/\rho N_A$   $\text{m}^3$ , giving an intermolecular distance of  $\bar{d} = \sqrt[3]{m_M/\rho N_A}$ . Thus  $\bar{d} = \sqrt[3]{28 \cdot 10^{-3}/800 \cdot 6.02 \cdot 10^{23}} = 0.387$  nm, about the same distance as the molecule ‘diameter’ we found earlier, as expected for a liquid.

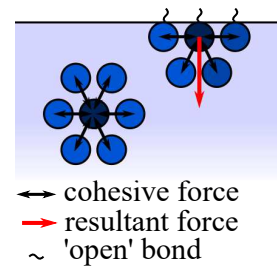
---

The net force is 0 for an intermolecular distance  $r_0 = \sqrt[6]{2}\sigma$ , while it becomes

rapidly strongly repulsive at shorter distance and attractive if the distance increases. This particular distance is actually the average molecules separation in liquid, while in gas this average distance is roughly 8 times larger. In fact a simplified view of a molecule in a fluid would be a hard sphere of diameter  $r_0$ , allowing to have simple illustration of the global arrangement of molecules in fluids as presented in the inset. The attractive part of the force is seen inside a fluid as a cohesive force that tries to keep its molecules together - a picture particularly true in liquid where the intermolecular distance is shorter and the force stronger. This cohesive force<sup>3</sup> is the cause of multiple phenomena in fluids like surface tension, capillarity, viscosity...

#### 4.3.3.1 Static properties

Using a naive approach suggested by the Lennard-Jones potential analysis we may sketch the forces between molecules in a fluid as sketched in the inset. This 'analysis' reveals a difference in the forces at play between the molecules in the bulk of the fluid and those at the surface in contact with vacuum or a gas. As we will see later, these differences could help us understand the origin of many properties of fluids.



- First, the interface exposes at the surface the cohesive force existing in the bulk and creates a tensioned like film, which is described as surface tension. It is this cohesive force that helps water strider walk at the surface of water, a liquid that has strong cohesive force as H<sub>2</sub>O molecules are polar and will create stronger bond (the H<sub>2</sub> bond) than what simple dispersive van der Waals forces can achieve.
- Then, we see that the symmetry around the surface molecules is broken resulting locally in an unbalance of force. The resultant force is normal to the surface and pulls the liquid molecule toward the center of the liquid body. Gradually, if there is no other forces, it will gather the liquid together until it assumes a spherical shape as the sphere keeps a stable shape under radial force. This property explains the characteristic spherical shape of droplets. At this stage, the force unbalance is still present and results in a rise of pressure inside the liquid.
- The interface also exposes a liquid surface with dangling van der Waals bonds ready to pair with external molecule and attract other fluid or solid. Actually, this would create adhesive forces but again other forces, like chemical, polar or ionic, may appear besides the van der Waals dispersive force. This force

<sup>3</sup>Note that in solids, cohesive force are usually much stronger than the one originating from the van der Waals forces, as they have other origins like covalent bonding, ionic bonding, metallic bonding...

will draw or repel liquid when they are in contact with solids and dictates the shape of droplet on flat surfaces.

Let's have now a look at this different effects in more details.

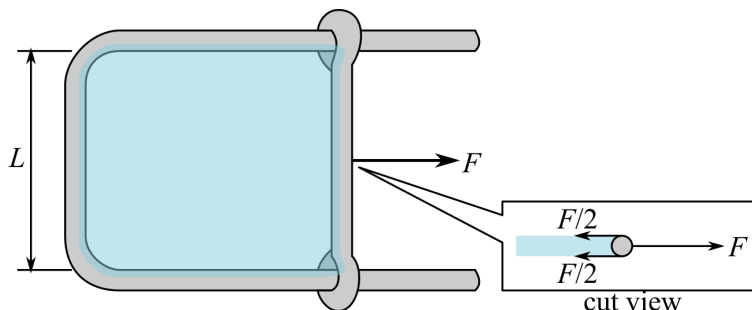


Figure 4.13: Measuring the surface tension with a U-shaped frame with a translating pin end (inset: cut-out view of the liquid in contact with the translating pin).

The surface tension is measured experimentally by forming a film of a certain liquid in a U-shape wire with a mobile pin, as shown in Figure 4.13. We observe that we need to apply a force  $F$  to maintain the mobile pin in place, demonstrating that an equivalent force is pulling the pin to the left. This force is due to the presence of the liquid film and linked to the liquid cohesive force. From the experiment we find that the surface tension is a force tangent to the surface but is not a spring force as  $F$  remains constant, independent of the pin translation distance. The force is proportional to the length  $L$  of the wet pin and we define the surface tension coefficient  $\sigma$  as:

$$F = 2\sigma L$$

the factor of 2 arising because of the two surfaces of the liquid film. The surface tension  $\sigma$  is thus a force per unit of length expressed in N/m.

We list in Table 4.7 the value of the surface tension coefficient for different fluids. Typical liquids have values between 10 mN/m and 80 mN/m. Actually, water is found to have the highest surface tension among common liquids, because the polar  $\text{H}_2\text{O}$  molecules attract each others creating a cohesive force much larger than the simple van der Waals dispersive force. Alcohols, oils, ethers and other organic solvents that are formed of non-polar molecules are in the lower region between 15 to 30 mN/m. Liquid metals have an order of magnitude higher surface tension.

As higher temperatures increase the disorder, it is found that they correspond to lower cohesive forces and surface tension – for example the surface tension of water drops to 59 mN/m at  $100^\circ\text{C}$ . Additionally, special chemical, called surfactants, do a much better job at reducing surface tension and for water essentially remove the effect of the  $\text{H}_2$  polar bond, bringing the surface tension down to 10-15 mN/m.

<b>Liquid</b>	<b>Liquid-gas surface tension (mN/m)</b>
Mercury	425
Sodium Chloride 6M	82
Water	72
Hydrochloric Acid (conc.)	70
Ethylene glycol	48
Isopropanol	23
Ethanol	22
Perfluorohexane	12

Table 4.7: Surface tension for selected fluids at 20°C (the surface tension will decrease with the temperature and with the presence of surfactants).

The surface tension can also be seen as the energy of a surface, the free surface energy. Actually in the experiment of Figure 4.13, we may pull the pin to the right and in doing so we will increase the surface of the wet film. The work produced for moving the pin by a distance  $\Delta x$  is  $W = \vec{F} \delta x = 2\sigma L \Delta x$ . This translation of the pin has actually increased the surface area by  $\Delta S = 2 \cdot \Delta x \cdot L$  (the 2 factor is again because of the two sides of the liquid film). The work produced is not lost and must appear in the new surface created as energy, whose density is given by :  $W/\Delta S = \sigma$ . Finally it means that the surface tension coefficient is also the energy density existing in the liquid film surface, and as such can be expressed in J/m<sup>2</sup>. This energy is called the free surface energy.

This energy helps give a quantitative view of the cohesive energy existing in the fluid. If we consider two planes of fluid in contact with each other, the van der Waals force between these two planes is expressed using  $A$  the Hamaker constant

$$F(r) = -\frac{A}{6\pi r^3} S$$

with  $S$  the surface area of the planes and  $r$  the separation distance<sup>4</sup>. To separate these two planes from contact where  $r = r_0$  to  $r = \infty$  we need a work of :

$$W = \int_{r_0}^{\infty} F(r) dr = -\frac{A}{12\pi r_0^2} S$$

This work is actually equal to the cohesive potential energy that was maintaining the two fluid planes in contact. From what we have seen above, the surface energy

<sup>4</sup>We have with interacting planes a factor of  $1/r^3$  in the expression of the force that is different from the  $1/r^7$  factor obtained earlier for individual molecules where induced dipoles are weaker.

created is  $2\sigma S$  (the factor of two arises because we have two planes now where we had a single interface), that is :

$$W = \Delta E_p = 2\sigma S$$

Thus the surface energy is given by :

$$\sigma = -\frac{A}{24\pi r_0^2}$$

Real free surfaces (characterized by free surface energy) are just possible for solids in vacuum and all other cases (gas/solid, solid/liquid, liquid/liquid) deal with interfaces between two phases and are characterized by interfacial energy. The interfacial energy is always a function of the two phases at the interface: solid / liquid, liquid / liquid, liquid / gas. For purely dispersive interaction between the two phases at the interface, the interfacial energy  $\sigma_{12}$  may be expressed as a function of the free energy of each phases:

$$\sigma_{12} \approx \sigma_1 + \sigma_2 - \sqrt{\sigma_1\sigma_2}$$

where  $\sigma_1$  and  $\sigma_2$  is the free surface energy of each phases. Contribution of the gas phase to the interfacial energy can be neglected in most cases, thus solid/gas (liquid/gas) surface energy is, respectively, almost like the solid (liquid) free surface energy.

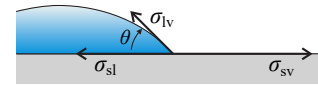
When the two surfaces are made of the same or similar materials, the energy that bind them is called the cohesive energy, but when we have different materials we talk about adhesive energy. Actually, the difference between the interfacial energy and the free surface energy of the two surfaces after separation is an adhesive energy  $W_A$  corresponding to the work that was needed to separate them (with a minus sign). It is expressed simply as :

$$W_A = \sigma_1 + \sigma_2 - \sigma_{12}$$

The dispersive van der Waals forces exists between any materials but other forces will also play a role in adhesion. The importance of cohesive and adhesive forces is best seen in 3 phases systems (gas, liquid, solid) with the spread of the liquid on the solid surface:

- If the cohesive force in the liquid is stronger than the adhesive force with the solid, the liquid forms beads on the solid surface
- If the adhesive force with the solid is stronger than the cohesive force in the liquid, the liquid spreads on the solid surface

This general issue is known as wetting. When a small lump of liquid is progressively poured on a solid surface it generally forms a drop whose shape after some time remains fixed. At that time, for the molecules lying on the stationary line interface between the three phases (gas, liquid and solid) the equilibrium of surface tension forces can be written using the Young equation as:



$$\sigma_{sl} + \sigma_{lg} \cos \theta = \sigma_{sg} \tag{4.2}$$

where  $\sigma_{sl}$ ,  $\sigma_{lg}$  and  $\sigma_{sg}$  are the surface tensions at the solid-liquid, the liquid-gas and the solid-gas interfaces respectively, and  $\theta$  is the contact angle.

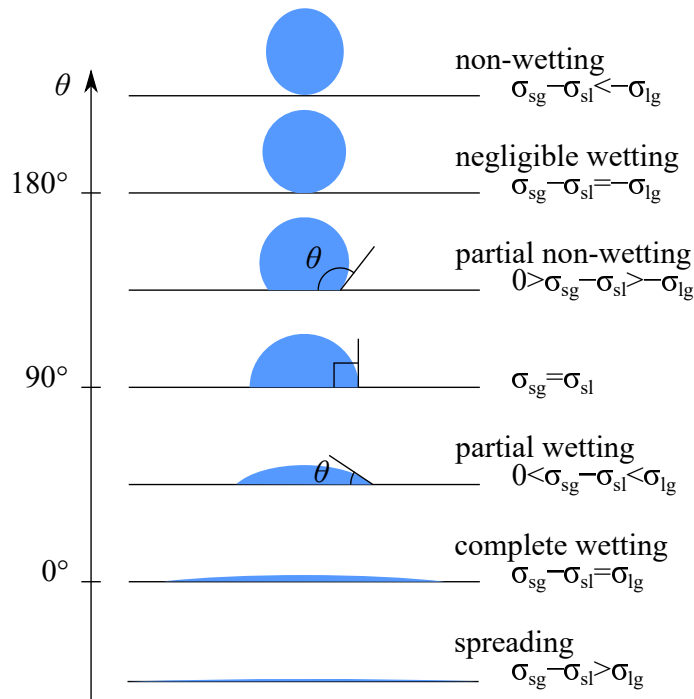


Figure 4.14: Wetting behaviour as a function of contact angle.

The contact angle is used in particular to distinguish between hydrophobic and hydrophilic surfaces, with the exact criterion being that the contact angle should be larger than  $90^\circ$  for the former and lower for the later as shown in Figure 4.14. The terms lyophilic and lyophobic are also used for the wetting behaviour of non-polar liquid (e.g. oil), that behaves differently from polar molecules as water when electrostatic forces are present at the solid surface. Moreover this surface property is often used in a less rigorous manner and describe relative value of different concepts as shown in Table 4.8

When the interface between two immiscible fluids (gas/liquid or liquid/liquid) shows a curved profile the force exerted toward the bulk of the liquid by the cohesive force at the liquid surface changes. Actually, as we see in Figure 4.15,



Parameters	Hydrophobic surface	Hydrophilic surface
Drop behavior		
Contact angle	high	low
Adhesiveness	poor	good
Wettability	poor	good
Solid surface free energy	low	high

Table 4.8: Properties of hydrophobic and hydrophilic surface.

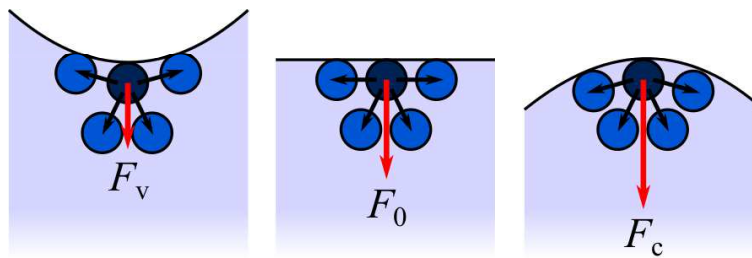


Figure 4.15: Resultant force due to cohesive energy between molecule at the surface of a liquid/gas interface with different radius of curvature.

when the surface is convex, the resulting force  $F_v$  is smaller than when the surface is flat ( $F_0$ ), and even smaller than when the surface is concave ( $F_c$ ). In practice, this force difference is expressed as a difference of pressure between the concave and convex side of a liquid/gas interface.

Actually if we consider a spherical droplet of radius  $r$  that changes radius by an amount  $\Delta r$ , its surface area  $A = 4\pi r^2$  change as :

$$\Delta A = 8\pi r \Delta r$$

resulting in an increase in surface energy of :

$$\Delta U = \sigma \Delta A = \sigma 8\pi r \Delta r$$

The mechanical work needed for this transformation is given by :

$$W = \vec{F} \cdot \vec{dl} = \Delta P S \Delta r = \Delta P 4\pi r^2 \Delta r$$

where  $\Delta P$  is the pressure difference between the inside and outside of the droplet.

Accordingly this difference of pressure  $\Delta P$  is given by :

$$W = \Delta U \quad (4.3)$$

$$\Rightarrow \Delta P 4\pi r^2 \Delta r = \sigma 8\pi r \Delta r \quad (4.4)$$

$$\Rightarrow \Delta P = \sigma \frac{2}{r} \quad (4.5)$$

We may generalize this expression for a more general curved interface where there are two principal radii of curvature  $R_1$  and  $R_2$ . This may describe in this way different rounded shapes like a saddle for example where  $R_1 > 0$  and  $R_2 < 0$  or a sphere where  $R_1 = R_2 = R$ . In this more general case we obtain the Young-Laplace equation:

$$p_{\text{concave}} - p_{\text{convex}} = \Delta p = \sigma_{\text{lg}} \left( \frac{1}{R_1} + \frac{1}{R_2} \right) \quad (4.6)$$

We verify that when  $R_1 = R_2 = R$  (spherical case) we get the previous expression, and we see that the pressure difference depends on the radius of curvature, and will change sign with it. We note that the smaller the radius of curvature, the larger the pressure difference. Actually it is this phenomena that cause the break-up of stream of water from a faucet: some initial irregularities in the diameter of the stream create zones with higher pressure (smaller diameter) and zones with lower pressure (larger diameter) and for a certain diameter these irregularities amplify with water flowing from high pressure zone to low pressure zone increasing the diameter difference and the pressure difference, ending up in the break-up of the stream. This phenomena is known as the Plateau-Rayleigh instability.

---

#### Example 4.3 Capillary pressure

---

FOR A SPHERICAL droplet, the two principal radii of the surface are equal  $R_1 = R_2 = R$  and we have  $\Delta p = 2\sigma_{\text{lg}}/R$ . Accordingly a droplet of water of 1 mm diameter in air will have an internal pressure higher than the external pressure by  $\Delta p = 2 \cdot 0.072/0.0005 = 288$  Pa that is about 0.003 atm, a very small value. Now if we consider a micro-bubble of air with a diameter of 10  $\mu\text{m}$  in water, we have  $\Delta p = 2 \cdot 0.072/5 \cdot 10^{-6} = 28800$  Pa that is about 0.3 atm, and if the bubble is in the sub-micrometer range the pressure difference will reach several atm.

We now consider a liquid jet sent from a high pressure orifice as can be used for cutting material. The jet will have a somewhat cylindrical shape. It means that it will have a finite radius of curvature in only one direction and for example  $R_1 = R$  across the cylinder and  $R_2 = \infty$  along the length. In that case the excess pressure inside the liquid jet will be given by  $\Delta p = \sigma_{\text{lg}}/R$ . As such, for a jet with the same diameter as the bubbles/droplets above, the pressure difference will be half the pressure found previously.

---

Using these two equations we can find the force exerted on a liquid in a circular micro-channel – also called a capillary. In the capillary, the adhesive or repulsive force appearing between the liquid and the solid curves the interface and give it



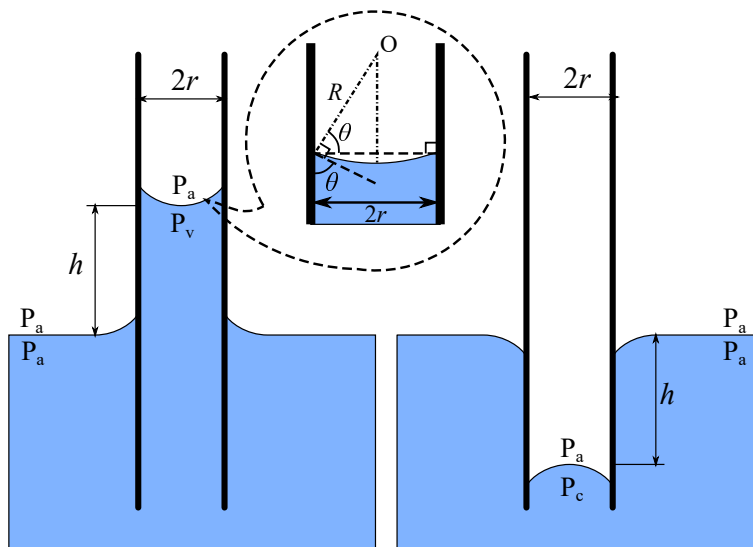


Figure 4.16: Liquid in wetting (left) and non-wetting (right) capillaries.

a spherical shape. Actually the radial symmetry of the capillary ensures that we have a single radius of curvature  $R_1 = R_2 = R$ . From simple geometry from the inset of Figure 4.16 we deduce that  $R = r/\cos\theta$  where  $r$  is the radius of the capillary. Then, the pressure drop at the liquid-gas interface simplifies to:

$$\Delta p = \frac{2\sigma_{lg}}{r/\cos\theta} \quad (4.7)$$

Depending on the contact angle of the liquid on the capillary wall we will observe two different behaviours:

- If the liquid wets the walls of the capillary ( $\theta < 90^\circ$ ) as seen on the left in Figure 4.16 it rises to a height  $h$ . Actually, the pressure below the meniscus  $P_v$  is lower than the atmospheric pressure and an equilibrium is achieved when the hydrostatic pressure ( $\Delta p = \rho gh$ ) caused by the weight of the water column will compensate this pressure loss.
- If the fluid does not wet the walls of the capillary ( $\theta > 90^\circ$ ) as seen on the right in Figure 4.16 it goes below the surface to a depth  $h$ . Actually the pressure below the meniscus  $P_c$  is higher than the atmospheric pressure and correspond to the pressure at a depth  $h$  in the liquid.

In both cases we have  $\frac{2\sigma_{lg}}{r/\cos\theta} = \rho gh$ , that is :

$$h = \frac{2\sigma_{lg}}{\rho gr/\cos\theta}$$

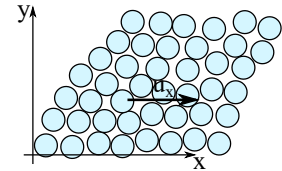
Capillarity can make efficient pumps in microsystems – as in trees where it allows water molecules to reach the leaves at a height of 50-60 m. Actually, when one

water molecule evaporates in a leaf, the capillary force pulls a neighbouring molecule up and the cohesive force in the liquid will pull molecules down the trunk to the roots. Capillarity makes it difficult to fill hydrophobic channels, as it tends to push the liquid back, or to empty hydrophilic ones, where it pulls the liquid inside. This effect is more pronounced for liquid with larger surface tension (Table 4.7), like water, which has one of the largest surface tension of common material due to the hydrogen bond between water molecules.

#### 4.3.3.2 Dynamic properties

In our naive view of the intermolecular forces in liquid, we saw that the main effect of these cohesive forces is to keep the fluid molecules in close proximity. As the molecules are set in collective motion, the molecules then collide into each others (clearly more so in liquid than in gas as the intermolecular distance is shorter in the former) giving rise to energy dissipation. This effect is macroscopically described as the viscosity that shows the resistance to flow of fluids.

In fact, Isaac Newton described flow of fluid as layers of fluid gliding on each other and postulated that the shear stress,  $\tau = F/S$  where  $F$  is the force needed to slide a layer of area  $S$ , is proportional to the velocity gradient in the direction perpendicular to the flow direction:



$$\tau = \eta \frac{\partial u_x}{\partial y}$$

where the proportionality coefficient  $\eta$  is the dynamic viscosity. This rule holds for so called Newtonian fluids (e.g. water, oil...), but there are many non-Newtonian fluids, particularly those containing particle (e.g. blood...), that won't follow this rule exactly. In this case, the viscosity is not a constant and depends on velocity. This so called dynamic viscosity  $\eta$  (or  $\mu$  in some books) is expressed in Pa/s (or in Poise, 1 P=0.1 Pa/s or more commonly in centipoise, 1 cP=1 mPa/s). When we are interested in ratio of friction to inertia force we often use the kinematic viscosity  $\nu = \eta/\rho$  expressed in m<sup>2</sup>/s (or in Stokes, 1 St=10<sup>-4</sup> m<sup>2</sup>/s) with  $\rho$  the density.

Viscosity for gas increases with temperature: for di-oxygen dynamic viscosity increases from 0.020 mPa/s at 20°C to 0.029 mPa/s at 200°C. The increased energy dissipation is linked to the increased number of collision as the molecule velocity increases. For liquids, viscosity decreases with higher temperature: for water dynamic viscosity drops from 1 mPa/s at 20°C to 0.28 mPa/s at 100°C. The reason seems to be in the thermal agitation driven increase in distance between the liquid layers that helps decrease the collision rate, and hence the viscosity. The viscosity varies in a large range over many orders of magnitude – tar, which is considered a fluid, is so viscous it takes years to form drops – and beside the easy to remember value for water ( $\eta = 1$  mPa/s), we may note the greater than 1 order of magnitude viscosity difference between liquid and gas.

<b>Fluid</b>	<b>Viscosity @ 20°C</b> (mPa·s)	<b>Fluid</b>	<b>Viscosity @ 20°C</b> (mPa·s)
honey	10 000	glycerin	1490
olive oil	84	blood	3-4
blood plasma	1.500	ethanol	1.200
water	1.002	benzene	0.652
acetone	0.32	oxygen	0.020
air	0.019	hydrogen	0.008

Table 4.9: Viscosity coefficient for different fluids at 20°C

In microfluidics fluid circulation may have different purposes: circulate species of interest in a carrier fluid (usually water) for biosensors, using flow forces to form droplet or bubbles in diphasic system, transport heat in cooling system...



Figure 4.17: Motion of particle in a fluid induced by only diffusion (left) or by only advection (right).

The issue of moving species of interest in fluidic systems is actually part of the more general mass transport phenomena, and even using simpler incompressible fluid, is rather complex. In fact there are two distinct phenomena that could produce mass transport, diffusion and advection (Figure 4.17). Diffusion is caused by a fundamental temperature dependent effect that spreads species inside the carrier fluid while advection<sup>5</sup> describes the transport of the species driven by the bulk motion of the fluid.

In diffusive flow, the species under motion are driven by collision with liquid molecules through Brownian (thermal) motion. If at small scale the motion is random, at larger scale it tends to spread the transported species uniformly, resulting globally in mass transport. Reaching the equilibrium state by diffusion takes long time and at large scale we usually see only the advection effect – but at micro-scale diffusion effect becomes significant. Thermodynamically, we can

<sup>5</sup>Advection is sometimes improperly called convection but convection is the transport (generally of heat) by the combined effect of diffusion and advection – that is what make sure the cold water in a pot placed on a furnace becomes (practically) uniformly hot.

say that through diffusion the system will move toward a state of maximum disorder, where the concentration of molecules is the same everywhere, to maximize its entropy. It is described by studying the evolution of the concentration of the

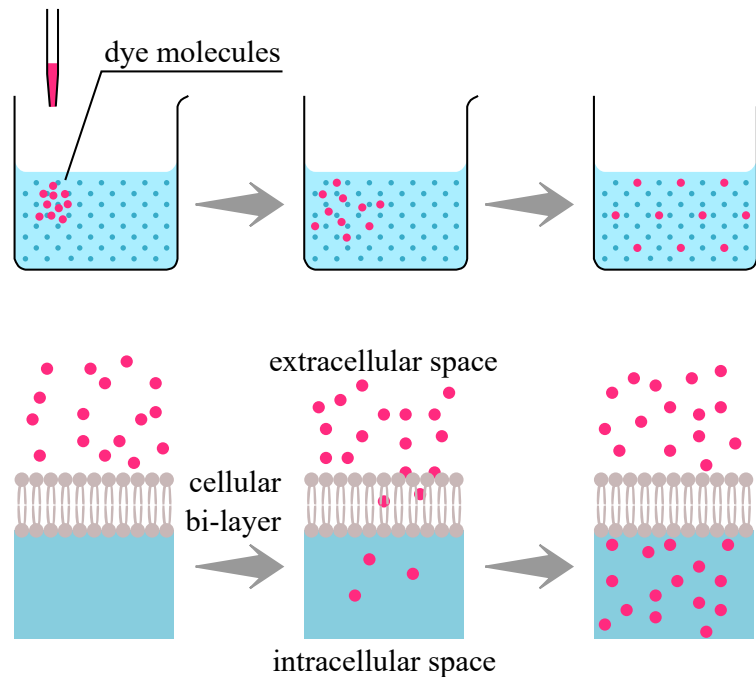


Figure 4.18: Diffusion of dye in a beaker (top) and amino acids at cellular boundary (bottom) representing, respectively, finite source and infinite source diffusion cases.

transported species. There are two main cases of practical interest that are shown in Figure 4.18:

- A drop diffusing in a reservoir (corresponding to the finite source case – e.g. drop of dye)
- The diffusion at the interface with a large reservoir (corresponding to the infinite source case – e.g. cell interface)

As we have seen in Section 3.5.1.2, the diffusive flux of particle  $j_d$  is proportional to the concentration gradient:

$$j_D = -D \overrightarrow{\text{grad}} C$$

where  $D$  is the diffusion coefficient in  $\text{m}^2/\text{s}$ . This coefficient varies with temperature, the viscosity of the liquid and the size of the diffusing molecules. In fact, we find that the larger the molecule is, the smaller the diffusion coefficient becomes, and for example we have:

- Water in water (“self-diffusion”) :  $D = 3.0 \cdot 10^{-9} \text{ m}^2/\text{s} @ 300\text{K}$

- Rhodamine-6G in water (fluorescent dye) :  $D = 0.3 \cdot 10^{-9} \text{ m}^2/\text{s}$  @ 300K
- Protein in water (IgG antibodies) :  $D = 0.05 \cdot 10^{-9} \text{ m}^2/\text{s}$  @ 300K

Moreover, we experimentally observe that with higher temperature the diffusion becomes faster. Actually it can be shown for spherical diffusing particle that:

$$D = \frac{k_B T}{6\pi\eta r_0}$$

with  $R$  the particle radius (we have  $R \approx r_0/2$  for molecules),  $T$  the absolute temperature in K and  $k_B = 1.38 \cdot 10^{-23} \text{ J/K}$  the Boltzmann constant.

To obtain the instantaneous amount of molecules flowing by diffusion per second for a certain concentration gradient we need to integrate the diffusive flux over a suitable cross-section. However the diffusion itself modifies the concentration gradient, and for obtaining the diffusion distance over a certain time we need to solve the equation given by Fick's second law

$$\frac{\partial C}{\partial t} = D\Delta C$$

By using proper boundary and initial conditions, the results derived in Section 3.5.1.2 show that for the two cases of interest (finite or infinite source) the diffusion length measured from the initial position (e.g., drop of dye or cell membrane) is given by:

$$d = k\sqrt{Dt}$$

where  $k$  is a constant depending on the surface concentration, the target concentration at the distance  $d$  and the type of source. It means that the diffusion distance increases as the square root of time, thus the velocity at which the diffusion front advances, tends to decrease over time as  $v_d = \partial d/\partial t = kD/2\sqrt{Dt}$ . Said in another way, the time it takes for the diffusion front to reach a certain distance  $d$  is given by  $t = d^2/k^2D$  – to reach 2 times further it will take 4 times longer. We understand that diffusion is not fast, particularly because typical values of  $D$  are small, and this physical phenomenon would only be useful at small scale<sup>6</sup>. For example, at small scale it is possible to use diffusion for mixing. Actually, mixing is normally obtained by ‘stirring’ — that is by creating turbulence. However, turbulence is hard to obtain in microfluidics and in that case diffusion would mix the fluids in an acceptable time. The effect can even be made faster by using multi-lamination: we divide the flow from a wide channel in  $n$  narrower channels, reducing the diffusion distance by the same factor. Accordingly, the mixing time is reduced by a factor  $1/n^2$  and may become really fast.

In advective flow the transported species move with the bulk of the fluid that flows with an external force (pressure, gravity, wall motion...). If the species are small enough, in steady state, after a transient state where they accelerate, their

---

<sup>6</sup>... or at high temperature.

motion is governed by the bulk motion of the carrying fluid. In general fluid flow is governed by the Navier-Stokes equation, a complex non-linear equation that can only be solved numerically for general configurations. However, in some practical cases at micro-scale this equation can be simplified and has even a few analytic solutions. The distinction between the cases is based on a series of dimensionless coefficients that compares different characteristics of the flow by computing ratio of fluid properties ( $u$  for the velocity of flow,  $c_s$  the speed of sound in the fluid,  $\rho$  the density of the fluid,  $\eta$  the dynamic viscosity,  $L$  a characteristic dimension of the system, usually its width).

- the Reynolds number ( $Re = \rho u L / \eta$ ) that corresponds to the ratio of inertial forces to viscous forces and measures how turbulent the flow is. At microscale with the characteristic length  $L \approx 100 \mu\text{m}$  and the typical flow velocity ( $u \approx 1 \text{ mm/s}$  to  $100 \text{ cm/s}$ ) we have  $Re < 1500$  the typical value that hints at the onset of turbulence. In that case there are no turbulence in the flow and we talk about laminar flow.
- the Mach number ( $Ma = u/c_s$ ) that corresponds to the ratio of the fluid velocity to the velocity of sound in that fluid and shows the flow compressibility. Actually, for  $Ma < 0.3$ , a gas can be considered as an incompressible fluid and use the same simplified equations as for liquids – for example in air the velocity of sound is approximately  $c_s = 340 \text{ m/s}$ , that is air can be considered an incompressible fluid for speed  $u < 100 \text{ m/s}$ .
- the Péclet number ( $Pe = Lu/D$ ) that corresponds to the ratio of the advective transport rate to the diffusive transport rate and shows for transported species if advective flow dominates ( $Pe \gg 1$ ) or if diffusive flow dominates ( $Pe \ll 1$ )
- the Knudsen number that we use at low pressure (Section 3.3) and essentially shows if we need to treat the flow using continuous mechanics or consider individual molecules.

For incompressible fluid with laminar flow, a condition often encountered in microfluidics, the Navier-Stokes equations can be simplified and becomes the Stokes equation:

$$\overrightarrow{\text{grad}} p = \eta \Delta \vec{u} + \vec{F} \quad (4.8)$$

where  $\vec{F}$  is a body force that applies everywhere on the liquid (like gravity<sup>7</sup>). Flow needs also to verify the mass conservation equation written as

$$\text{div } \vec{u} = 0$$

and expressing that the net flow at any point is 0 (i.e. as much fluid entering than leaving).

---

<sup>7</sup>Actually gravity does not usually applies in microfluidics as most circuit are planar. Even when it is not the case, gravity forces can be neglected with respect to other forces like capillary forces, pressure...

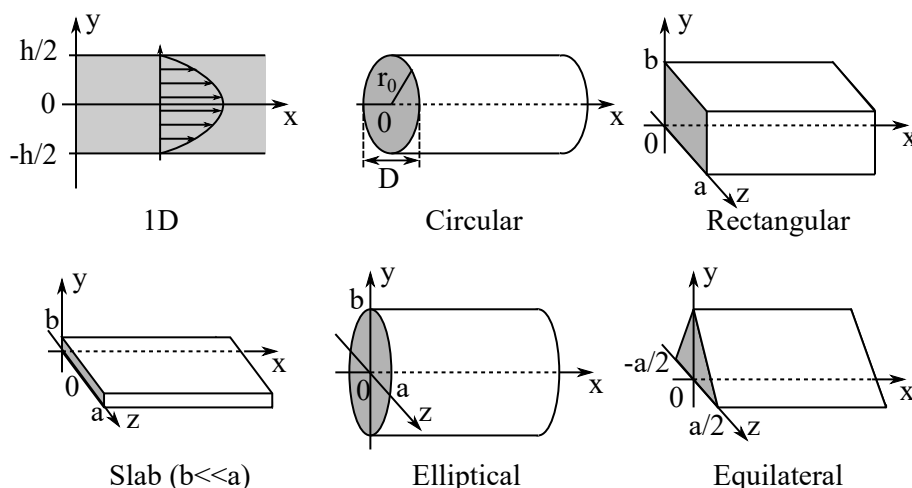


Figure 4.19: Typical channels used in microfluidics where Poiseuille flow has closed-form solution.

Within this limit, the steady flow induced by a difference of pressure (called a Poiseuille's flow) can be obtained analytically for some typical geometries relevant to microchannels as shown in Fig. 4.19.

Before we embark to detail these solutions, we have to stress that they only describe flow in a channel of constant cross-section, far from the entrance. In fact, it takes time for these flow to fully establish and an entry length is defined that ensure that past this distance, the flow is fully developed. For laminar flow, the entry length is given by:

$$L_h \approx 0.05 Re D \quad (4.9)$$

where  $D$  is the width of the channel and  $Re$  the Reynold's number. For typical microchannels where  $Re \approx 1$ , this distance represents about 5% of the channel width, but for high velocity  $\bar{u} \approx 50\text{cm/s}$ , the entry length may represent 2 to 3 times the channel width. Within this distance the velocity profile of the flow gradually changes from the profile at the entrance, to the fully developed profile corresponding to the solution of the Stokes equation.

For better showing the approach in using Stokes equation, we will consider the 2D channel (with 1D cross-section), without body force and start by writing the equation in Cartesian coordinates as:

$$\begin{cases} \frac{\partial p}{\partial x} = \eta \left( \frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2} \right) \\ \frac{\partial p}{\partial y} = \eta \left( \frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2} \right) \end{cases}$$

The pressure is applied uniformly across the channel height, thus  $\partial p / \partial y = 0$  and we consider steady state in the channel thus we have  $\partial^2 u / \partial x^2 = 0$ , simplifying the

system of equation to :

$$\begin{cases} \frac{dp}{dx} = \eta \frac{\partial^2 u_x}{\partial y^2} \\ 0 = \eta \frac{\partial^2 u_y}{\partial y^2} \end{cases}$$

We integrate once w.r.t.  $y$  giving:

$$\begin{cases} \frac{dp}{dx} y = \eta \frac{\partial u_x}{\partial y} + A \\ B = \eta \frac{\partial u_y}{\partial y} \end{cases}$$

For symmetry reason we see that the velocity  $\vec{u}$  has to be maximum in the center of the channel at  $y = 0$ , thus  $\frac{\partial u}{\partial y}(0) = 0$  and we have in this point:

$$\begin{cases} 0 = 0 + A \\ B = 0 \end{cases}$$

and  $A = B = 0$ . We integrate a second time w. r. t.  $y$ :

$$\begin{cases} \frac{1}{2} \frac{dp}{dx} y^2 = \eta u_x + C \\ D = \eta u_y \end{cases}$$

The second equation shows that  $u_y$  is constant in the channel. However, at the channel wall in  $y = \pm h/2$ , the liquid molecules adhere to the solid and their velocity is null  $\vec{u} = 0$ . Thus we have there  $u_y = 0$ , meaning that  $D = 0$  and that  $u_y = 0$  everywhere in the channel.

Placing ourselves on the wall again, the first equation becomes:

$$\frac{1}{2} \frac{dp}{dx} (h/2)^2 = 0 + C$$

thus  $C = \frac{1}{2} \frac{dp}{dx} (h/2)^2$ . Finally the velocity field can be written as:

$$\begin{cases} u_x = \frac{1}{2\eta} \frac{dp}{dx} (y^2 - (h/2)^2) \\ u_y = 0 \end{cases}$$

We see that the flow is occurring only along the  $x$  direction ( $u_y = 0$ ) and that the velocity profile across the channel is parabolic, which is a clear signature of Poiseuille's flow. In this channel the maximum velocity is at the center in  $y = 0$  and is expressed as:

$$u_{\max} = -\frac{dp}{dx} \frac{h^2}{8\eta}$$

Moreover the average velocity is given by:

$$\bar{u} = \frac{1}{h} \int_{-h/2}^{+h/2} \frac{1}{2\eta} \frac{dp}{dx} (y^2 - (h/2)^2) dy = -\frac{dp}{dx} \frac{h^2}{12\eta}$$



and we have  $\bar{u} = \frac{2}{3}u_{\max}$ .

In fact, it is not always necessary to know the velocity precisely at any point across the channel and often it is enough to obtain the average velocity  $\bar{u}$  or the flow rate. This last quantity comes in two flavours which should not be mixed: the volumetric flow rate (noted  $\dot{V}$  or  $Q$ ), which represents the *volume* of fluid passing across a certain section of channel over a unit time expressed in  $\text{m}^3/\text{s}$  or the mass flow rate (noted  $\dot{m}$ ), which represent the *mass* of fluid passing across a certain section of channel over a unit time expressed in  $\text{kg}/\text{s}$ . The volumetric flow rates is related to the average fluid velocity by:

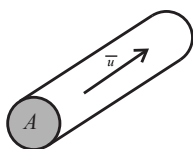
$$\dot{V} = \bar{u}A$$

where  $A$  is the channel cross-section area. The mass flow rate for incompressible fluids is accordingly given by:

$$\dot{m} = \rho\bar{u}A$$

where  $\rho$  the fluid density. For our 1D channel the volumetric flow rate is given by:

$$\dot{V} = \bar{u} \cdot h = -\frac{dp}{dx} \frac{h^3}{12\eta}$$



For incompressible fluids, the difference between the two flow rates is not too significant ( $\rho$  product) but for compressible fluids the distinction is important as the density may vary with time. In fact we have:

$$\dot{m} = \frac{dm}{dt} = \frac{d\rho V}{dt} = \rho \frac{dV}{dt} + V \frac{d\rho}{dt}$$

with  $\rho$  the fluid density. It is clear that if the fluid density does not change we have simply:

$$\dot{m} = \rho \frac{dV}{dt} = \rho \dot{V}$$

but the  $V \frac{d\rho}{dt}$  term remains if compressibility has to be taken into account.

From a physics point of view the mass flow rate is more precise as it defines a quantity of matter unambiguously (mass relates to number of molecules) but customary usage makes the volumetric flow rate more prevalent, even for (compressible) gas. This last case may create ambiguity<sup>8</sup> as it generally requires simultaneously the knowledge of temperature and pressure to ascertain the quantity of matter in the gas volume. The notation  $Q$  used for the volumetric flow rate adds also to the ambiguity as it seems to imply considering a certain Quantity of fluid... whereas it is merely a volume<sup>9</sup>.

<sup>8</sup>It is a similar reason that has compelled us to use mass flow rate for study of package leak in section 5.2.2, contrary to what is done in the standard literature.

<sup>9</sup>The parallel with total charge, also noted  $Q$ , and current in electricity is probably not fortuitous, with  $I = \dot{Q}$ , but there, the charge defines unambiguously a quantity of matter.

Cross-section	Fluid velocity profile, mean velocity and volumetric flow rate
1D	$u_x(y) = -\frac{dp}{dx} \frac{(h/2)^2 - y^2}{2\eta}$ , $\bar{u} = \frac{2}{3}u_{\max}$ and $\dot{V} = -\frac{dp}{dx} \frac{h^3}{12\eta}$
Slab for $a \gg b$	$u(z, y) = -\frac{dp}{dx} \frac{(b/2)^2 - y^2}{2\eta}$ , $\bar{u} = \frac{2}{3}u_{\max}$ and $\dot{V} = -\frac{dp}{dx} \frac{ab^3}{12\eta}$
Circular	$u_x(r) = -\frac{dp}{dx} \frac{r_0^2 - r^2}{4\eta}$ , $\bar{u} = 0.5u_{\max}$ and $\dot{V} = -\frac{dp}{dx} \frac{\pi r_0^4}{8\eta}$
Elliptic	$u(z, y) = -\frac{dp}{dx} \frac{a^2 b^2}{2\eta(a^2 + b^2)} \left(1 - \frac{z^2}{a^2} - \frac{y^2}{b^2}\right)$ and $\dot{V} = -\frac{dp}{dx} \frac{\pi a^3 b^3}{4\eta(a^2 + b^2)}$
Rectangular original for $a > b$	$u(z, y) = -\frac{dp}{dx} \frac{16a^2}{\pi^3 \eta} \sum_{n=1,3,\dots}^{\infty} (-1)^{\frac{n-1}{2}} \left(1 - \frac{\cosh(n\pi z/2a)}{\cosh(n\pi b/2a)}\right) \frac{\cos(n\pi y/2a)}{n^3}$ and $\dot{V} = -\frac{dp}{dx} ab \frac{4a^2}{3\eta} \left(1 - \frac{192a}{\pi^5 b} \sum_{n=1,3,\dots}^{\infty} \frac{\tanh(n\pi b/2a)}{n^5}\right)$
Rectangular improved for $a \geq b$	$u(z, y) = -\frac{dp}{dx} \frac{16b^2}{\pi^4 \eta} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{\sin(n\pi z/a) \sin(m\pi y/b)}{nm((b/a)^2 n^2 + m^2)}$ and $\dot{V} = -\frac{dp}{dx} ab \frac{b^2}{\eta} \frac{64}{\pi^6} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}$
Square	$u(z, y) = -\frac{dp}{dx} \frac{16a^2}{\pi^4 \eta} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{\sin(n\pi z/a) \sin(m\pi y/a)}{nm(n^2 + m^2)}$ and $\dot{V} = -\frac{dp}{dx} \frac{a^4}{28.45\eta}$
Triangular equilateral	$u(y, z) = -\frac{dp}{dx} \frac{\sqrt{3}}{2a\eta} y(z + y/\sqrt{3} - a/2)(z - y/\sqrt{3} + a/2)$ and $\dot{V} = -\frac{dp}{dx} \frac{\sqrt{3}a^4}{320\eta}$

Table 4.10: Analytic expressions of fluid velocity and volumetric flow rate in channels with 1D, slab, circular, elliptic, rectangular (original solution from Joseph Boussinesq and improved solution [19]), square and equilateral triangular cross-sections (refer to Figure 4.19 for coordinate and geometric parameters explanation).

The Stokes equation (4.8) that we have used for solving the hydrodynamic 1D problem has an infinite number of closed form exact solutions but only a few correspond to geometries useful for the designer. The most interesting results have been summarized in Table 4.10, and the derivation of the circular channel case is the object of Problem 10. For example, for the pressure driven flow in a circular channel of diameter  $D = 2r_0$ , length  $L$  and with a difference of pressure  $\Delta p$ , also called a Poiseuille's flow, we have:

$$\dot{V} = \frac{\Delta p \pi D^4}{L 128\eta} \quad (4.10)$$

thus giving an average velocity of :

$$\bar{u}_{\text{pf}} = \frac{\dot{V}}{\pi D^2/4} = \frac{\Delta p D^2}{L 32\eta}$$

A phenomenological equation may be obtained using the Darcy-Weissbach approach, that is often used for channel cross-sections for which there is no known solutions. One of the form of this equation is:

$$\dot{V} = \frac{2AD_h^2}{Re f \eta L} \Delta p \quad (4.11)$$

and thus:

$$\bar{u} = \frac{2D_h^2}{Re f \eta L} \Delta p \quad (4.12)$$

where  $Re f$  is the product of the Reynold's number ( $Re$ ) and  $f$  the Darcy's friction factor,  $A$  the channel cross-section area and  $D_h$  the hydraulic diameter. This parameter is given by:

$$D_h = \frac{4A}{P_{\text{wet}}} \quad (4.13)$$

where  $P_{\text{wet}}$  is the wet perimeter, that is, the length of the channel perimeter in direct contact with the fluid. We note that for a circular channel of diameter  $D$  we have  $D_h = 4A/P_{\text{wet}} = 4\pi(D/2)^2/2\pi(D/2) = D$ , as we would expect, and for a square channel of side  $a$ ,  $D_h = 4a^2/4a = a$ .

The  $Re f$  product can be computed from first principle for circular channel and if we compare equation (4.11) with equation (4.10) for the flow rate, we see that we have in this case  $Re f = 64$ . The accuracy of this approximation is low and care should be taken when hydraulic diameters are used for profiles diverging greatly from circular profiles, and more exact formulas from Table 4.10 should be used. In fact the  $Re f$  product is usually taken as an adjustment parameter and tables give its value for different channel sections differing from the circular section. For example for a rectangular section this parameter varies from 96 for a slab profile (a wide 1D profile) to 56.9 for a square profile, showing that the equation with a constant  $Re f = 64$  is subject to a maximum error of  $(96 - 64)/64 = 50\%$  when it is used for square cross-section...

Another form of generalized equation has been proposed where the hydraulic diameter is used to define an equivalent circular channel and get rid of  $A$  the channel cross-section area:

$$\dot{V} = \frac{\pi D_h^4}{2Re f \eta L} \Delta p \quad (4.14)$$

This equation has the advantage to show more clearly the  $D^4$  dependency of the flow rate but proves to be even less precise than the form above.

A closer look at the flow rate equation, reveals that using analogies (cf. Section 2.4.4) we can write it in a way that is essentially equivalent to Ohm's law, with the effort variable  $U$  being the pressure difference  $\Delta P$  and the flow variable  $I$  being the mass flow rate  $\dot{m}$ . We note here that we prefer to use the mass flow rate  $\dot{m}$  over the volumetric flow rate  $\dot{V}$ , as we discussed it earlier because it represent better the flow of a measurable quantity (extensive variable), mass, which is akin to charge in electricity. In fact we have:

$$\dot{m} = \rho \bar{u}_{pf} \cdot \pi(D/2)^2 = \frac{\pi D^4}{2Re f \nu L} \Delta p$$

where we use the kinematic viscosity  $\nu = \eta/\rho$ , using  $\rho$  the density of the fluid. We may write this equation as:

$$\Delta p = \frac{2Re f \nu L}{\pi D_h^4} \dot{m}$$

allowing us to define the hydrodynamic resistance,  $R_h$  given by:

$$R_h = \frac{2Re f \nu L}{\pi D_h^4}$$

This expression gives general geometry dependance and more detailed expressions are given in Table 4.11 for other profiles with analytic solutions. We see that the hydrodynamic resistance increases very rapidly with the decrease in diameter<sup>10</sup> and even if microchannel lengths are very short, the hydrodynamic resistance is significant at micro-scale and need to be taken into account in microfluidics circuit designs. Using this analogy, a network of channels would be represented by an equivalent network of resistance, and solving for the branch currents (node voltages) will solve for the flow rates (pressures), respectively.

We find also fluidic equivalence to other circuit elements, like fluidic capacitor or fluidic inductor. The hydrodynamic capacitance, with circuit constitutive equation given by  $i = C dv/dt$ , may thus be described by a pressure dependent mass flow :

$$\dot{m} = C_h \frac{dP}{dt}$$

which typically may arise in microfluidic circuits for two reasons:

---

<sup>10</sup>The analogy with Ohm's law does not hold for the definition of the resistance as we have  $R = \rho L/S$  in electricity, whereas in fluidics, if it is also proportional to the length of the channel  $L$ , it varies roughly at the inverse of the *square* of the section area  $1/S^2$

Cross-section	Hydrodynamic resistance
1D	$R = \frac{12\eta L}{h^3}$
Slab for $a \gg b$	$R = \frac{12\nu L}{ab^3}$
Circular	$R = \frac{8\nu L}{\pi r_0^4}$
Elliptic	$R = \frac{4\nu(a^2+b^2)L}{\pi a^3 b^3}$
Rectangular original for $a > b$	$R = \frac{3\nu L}{4a^3 b} \frac{1}{\left(1 - \frac{192a}{\pi^5 b} \sum_{n=1,3,\dots}^{\infty} \frac{\tanh(n\pi b/2a)}{n^5}\right)}$
Rectangular improved for $a \geq b$	$R = \frac{\nu L}{ab^3} \frac{\pi^6}{64 \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}}$
Square	$R = \frac{28.45\nu L}{a^4}$
Triangular equilateral	$R = \frac{320\nu L}{\sqrt{3}a^4}$

Table 4.11: Hydrodynamic resistance for laminar flow in channels of length  $L$  with the cross-sections from Figure 4.19 ( $R$  is given for mass flow-rate, use  $\eta$  instead of  $\nu$  for volumetric flow-rate).

- with incompressible fluids, if the channel or chamber walls are deformable, their volume  $V$  will change with pressure
- with compressible fluid, the density of the fluid will be changing with pressure

In fact we have here the two sources that were discussed in the definition of the mass flow rate

$$\dot{m} = \frac{d\rho V}{dt} = \rho \frac{dV}{dt} + V \frac{d\rho}{dt}$$

and they may obviously be mixed in case of compressible fluids in channel/chamber with deformable walls.

Considering the first term we may write:

$$\dot{m}_V = \rho \frac{dV}{dt} = \rho \frac{dV}{dP} \frac{dP}{dt}$$

thus the hydrodynamic capacitance due to deformable chamber/channel is written as:

$$C_V = \rho \frac{dV}{dP}$$

and if we consider the volume of the chamber given by  $V = V_0 + c(P - P_0)$  where we introduce the compliance  $c$  of the chamber,  $P$  the pressure inside the chamber and  $P_0$  the pressure outside, we have simply  $dV/dP = c$  and:

$$C_V = \rho c$$

Considering the compressibility issue, we assume we have a chamber of fixed volume  $V_0$  and the fluid is compressed adiabatically (does not change temperature during compression) by increasing the pressure  $P$ . We may write:

$$\dot{m}_\rho = V \frac{d\rho}{dt} = V \frac{d\rho}{dP} \frac{dP}{dt}$$

The gas compressibility is given by  $\kappa = \frac{-1}{V} \frac{dV}{dP}$ , which is expressed in terms of density by using the conservation of the mass  $m_0 = V\rho$ . We have  $d\rho = \frac{-m_0}{V^2} dV$  thus  $\frac{-1}{V} dV = \frac{V}{m_0} d\rho$  and  $\kappa = \frac{V}{m_0} \frac{d\rho}{dP}$  giving finally:

$$\dot{m}_\rho = \kappa m_0 \frac{dP}{dt}$$

thus the compressibility component of the hydrodynamic capacitance is given by:

$$C_\rho = \kappa m_0$$

The fluidic inductor is characterized by an hydrodynamic inductance that follows the electric inductor relationship  $v = L \frac{di}{dt}$ , and is thus described by a pressure depending on the variation of the mass flow-rate. Most of the time this is simply representing the inertia of the fluid, or its resistance to be accelerated. For example if we consider a straight channel of cross-section area  $A$  and length  $l$  with fluid at rest, that is subjected to a sudden increase of pressure at its entrance, we find that the hydrodynamic inductance is given by:

$$L_h = \frac{l}{A}$$

which will combine with the hydrodynamic resistance of the channel  $R_h$  to delay the establishment of a steady flow-rate in the channel, with a time constant given by  $\tau = R_h/L_h$ . It could be noted that when there is a sudden change of pressure, the pressure loss is first fully attributable to the inertia loss and then, as the fluid accelerates, it becomes more and more due to the viscous loss (that is given by  $R_h$ ), that becomes the only cause of loss when we reach a constant velocity. For typical channels in microfluidics, the time constant may be in the order of 20 ms.

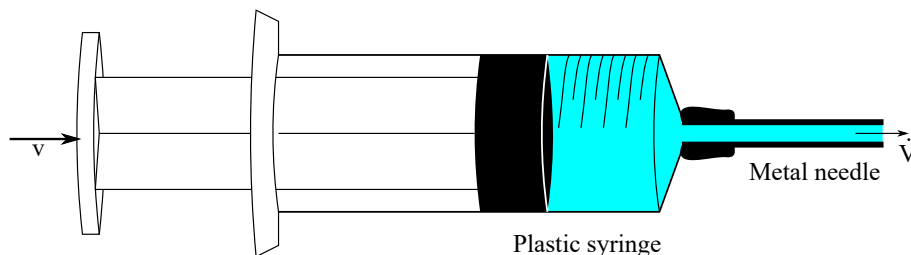
Before we close this analogy approach it is important to not forget the last elements of an electric circuit: the sources. In fact in fluidics as in electricity the sources comes in two flavors:

- Equivalent current sources are created by device imposing a flow-rate to the fluid, typically, a syringe pump.
- Equivalent voltage source are created by device imposing a pressure at one place in the circuit, typically, a pressure pump.

The last step required before the circuit variables could be solved is the connection of the different elements following the rules described in Section 2.4.4.

**Example 4.4** Flow from a syringe

WE CONSIDER THE INJECTION OF A LIQUID in a microfluidic channel using the syringe depicted in the Figure. The plastic 10 ml syringe (inner diameter  $\phi_s = 14.5$  mm) is placed in a syringe pump that pushes the plunger at a velocity  $v$ . The 32G stainless steel needle has an inner diameter of  $\phi_n = 108$   $\mu\text{m}$  and has a length  $L = 25$  mm. We will now describe the evolution of the flow rate of water at the needle output  $\dot{V}$ , if we start from an initially static syringe pump (that is  $v = 0$ ).



The first step is to decide how to model our syringe/needle system. It seems convenient to use a system approach and we choose a reduced order lumped element model based on a circuit approximation.

We need then to divide the complete system in connected subsystems that would be represented by circuit elements. We decide here to divide the system in : a moving plunger, a barrel and a needle. We will now model each of these subsystems as circuit elements, resistor, capacitor, inductor and source.

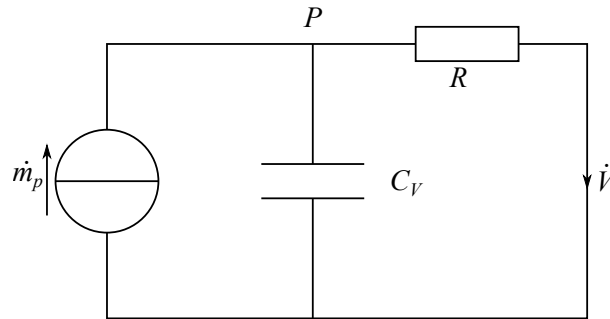
The plunger constant velocity  $v$  will impose a flow-rate at the entrance of the system, that is represented, from the analogies chosen, by a source of current. In fact we have a volumetric flow-rate given by  $\dot{m}_p = \rho v A$  where  $A = \pi(\phi_s/2)^2$  is the surface area of the plunger tip pushing the water and  $\rho = 1000$   $\text{kg}/\text{m}^3$  for water.. The barrel of the syringe may be considered as a channel of large width, but because it is made of plastic it may also expand if the water pressure is exceeding the outside atmospheric pressure. In fact this barrel should be represented by a resistance, to take into account the viscous loss in the water flow, a capacitor, to take into account the deformability of the barrel, and an inductor, to take into account the inertia of the water as it is accelerated from its rest position. As we want to establish the most important features of the flow, we will in a first approximation neglect the inductance (short time constant), and also neglect the viscous loss in the barrel as the needle is more than 100 times narrower than the barrel and will thus represent a much larger resistance in the system. Accordingly the barrel is represented by a simple capacitor, whose value is obtained from the hydraulic compliance of the plastic barrel  $c$  as  $C_V = \rho c$ .

The compliance  $c$  may be estimated by using the variation of radius of a cylinder under pressure  $\Delta r = pr^2/Et$  where  $t$  is the wall thickness and  $E$  the Young's modulus of the material. We obtain  $c = dV/dp = 2\pi r^3 l/Et$  and get  $c \approx 7 \cdot 10^{-14}$   $\text{m}^3/\text{Pa}$ , with  $E = 2$  GPa,  $t = 1$  mm,  $l = 60$  mm,  $r = 7.25$  mm.

**Example 4.4** Flow from a syringe (continued)

The steel needle is considered as a narrow channel, but because it is made of steel we will consider that its deformability is low and won't change significantly of dimension when the internal pressure varies. Accordingly (and we also neglect the inertia loss) the needle becomes a simple resistor, whose equivalent resistance is obtained from the formula in Table 4.10, as  $R = \frac{128\nu L}{\pi\phi_n^4}$ . We notice that  $\nu = \eta/\rho$  is the kinematic viscosity, and for water it is  $\nu = 10^{-3}/1000 = 10^{-6}$  m<sup>2</sup>/s.

Having determined the different circuit elements they need to be connected together. In fact the plunger tip, the inside of the barrel and the needle input are at the same pressure (we neglected pressure drop along the barrel) and accordingly should share the same node (with potential  $P$ ). The outside of the barrel and the needle output are also at the same reference pressure, and share a single node. The resulting circuit is thus the following:



From this circuit, we may easily write the governing equation for the flow-rate at the output of the needle, by establishing the circuit transfer function. We have:

$$\rho\dot{V} = \frac{P}{R} = \frac{\dot{m}_p \frac{1}{1/R + C_V s}}{R} = \frac{\dot{m}_p}{1 + RC_V s}$$

We recognize immediately a first order system, and referring to 2.11 we know that the response to a step input (a sudden start of the plunger) will be of the form :

$$\dot{V}(t) = vAu(t)(1 - e^{-t/RC_V})$$

It means that it will take  $3RC_V$  s for the needle output flow rate to reach 90% of what is imposed by the plunger of the syringe ( $vA$ ).

With the dimension given, we find  $3RC_V = 3 \frac{128\nu L}{\pi\phi_n^4} \rho c = 3 \frac{128 \cdot 10^{-6} \cdot 25 \cdot 10^{-3}}{\pi(108 \cdot 10^{-6})^4} 1000 \cdot 7 \cdot 10^{-14} \approx 15$  s. The time is reasonable but it should not be overlooked when doing experiment where we vary the flow rate: we have to wait about 30 s to make the measurement after setting a new flow rate.

Of course if we use a glass syringe, which has a Young's modulus about 50 times larger,  $c$  and this time will be divided by 50. Decreasing the volume of the syringe (1 ml) will also have an important benefit on this aspect.



**Example 4.4** Flow from a syringe (end)

An additional possibility is to use a larger needle, and by multiplying its diameter by 2 we already gain a factor of 16 thanks to the power of 4 of the needle diameter, allowing to reach the equilibrium very fast. However, if the channel diameter decreases by 2 (goes to 50  $\mu\text{m}$ ), then the time increases by a factor of 16 and the 15 s becomes 4 min.

Note that before it reaches the microchannel (here the needle) the syringe is generally connected to a relatively long tube made of soft material (PTFE, FEP...). But if we remain reasonable on the tube length (1 m) with 1/16" outer diameter, because of its small internal diameter (for example, 250  $\mu\text{m}$ ), the compliance of the tube is much smaller than the compliance of syringe barrel shown here, and we should be more worried by the compressibility of bubbles trapped in the tube that behave also like capacitors.

In conclusion, fast microfluidic circuits (particularly connected to high hydrodynamic resistance microchannels) should take care of this issue... or use a pressure source instead of a syringe pump (but that poses other problems !).

## 4.4 Sensor technology

Sensing is certainly a quality that we associate with living being. A stone does not sense, but can a silicon circuit do it? Of course, the answer is yes, and MEMS have increased tremendously the number of physical parameters that are sensed by silicon.

Sensing can be formally defined by the ability to transform any form of energy present in the environment into energy inside a system. An example will be to convert the air temperature (heat energy) to an electrical signal by using a thermocouple. At the heart of the sensor is the ability to perform the energy transformation, a process usually called transduction. MEMS sensors ability to measure different parameters as pressure, acceleration, magnetic field, force, chemical concentration, etc is actually based on a limited number of transduction mechanisms compatible with miniaturization : piezoresistive, capacitive, piezoelectric, and in fewer cases, inductive. In general this primary transduction mechanism won't directly produce a signal in a form suitable for further processing in an instrumentation chain. In fact, 95% of the time we want to obtain the signal as a variation of voltage, and we will need some signal conditioning to reach that goal. In Table 4.12 we list the different sensing mechanisms with the corresponding primary form of the signal and typical circuits used for obtaining voltage variation.

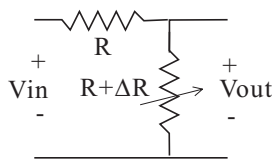
### 4.4.1 Piezoresistive sensing

The oldest MEMS sensor that gained huge popularity was the pressure sensor and it was based on the piezoresistive effect. Piezoresistivity can be described by the change of resistance of a material when it is submitted to stress. This effect is

Type	Measurand	Primary signal	Conditioning circuit
Piezoresistive	Stress	Resistance	Potentiometric, Bridge
Capacitive	Deformation	Capacitance	Bridge
	Permittivity		Frequency converter
Piezoelectric	Stress	Charge	C-V converter
Inductive	Deformation	Reluctance	Bridge
			Frequency converter

Table 4.12: Sensing principles and conditioning.

known since the 19th century in metals, but it was only in the mid 1950s that it was recognized that semiconductor and particularly silicon had huge piezoresistive coefficient compared to metal[4]. The MEMS designer will then create piezoresistors by doping locally silicon and place them where the stress variation is maximal, for example, at the edge of a membrane in pressure sensor. Then by measuring their resistance change he will be able to infer the stress which is related to the deformation.



For converting the resistance change  $\Delta R$  in a voltage, the potentiometric (voltage divider) configuration is the simplest, and the voltage at the output is given by :

$$V_{\text{out}} = \frac{R + \Delta R}{2R + \Delta R} V_{\text{in}} \approx \frac{V_{\text{in}}}{2} + \frac{\Delta R}{2R} V_{\text{in}}$$

However, as we see, this configuration suffers from a strong offset ( $\frac{V_{\text{in}}}{2}$ ), complicating the signal processing operation, and a relatively low sensitivity ( $s = dV_{\text{out}}/d\Delta R = \frac{V_{\text{in}}}{2R}$ ). A better configuration is based on the Wheatstone bridge circuit (Figure 4.20) that completely suppresses the offset and reach a better sensitivity in some configurations. The suppression of offset is obtained by balancing the bridge, that is obtaining a null output voltage when the resistor has its nominal value. This condition is obtained when the resistors in each branch of the bridge have values verifying  $R_1 R_3 = R_2 R_4$  – which is automatically reached if the four resistors are the same and  $R_1 = R_3 = R_2 = R_4 = R$ . From there, it is simple to show that if there is one single variable resistor in the balanced bridge and if  $\Delta R \ll R$  then

$$V_{\text{out}} \approx \frac{V_{\text{in}}}{4R} \Delta R.$$

We see here that we have suppressed the offset completely allowing easy amplification further down the chain. Moreover, we can increase the sensitivity to make it surpass the earlier voltage divider configuration. By positioning the variable resistors with the configuration shown on the right (where there are four variable

resistors and where the change of resistance induced by the measurand on two of the resistors is opposite to the change induced on the two other but with the same magnitude), then we exactly have

$$V_{\text{out}} = \frac{V_{\text{in}}}{R} \Delta R$$

and the sensitivity of the bridge configuration increases fourfold

$$s = \frac{V_{\text{in}}}{R}.$$

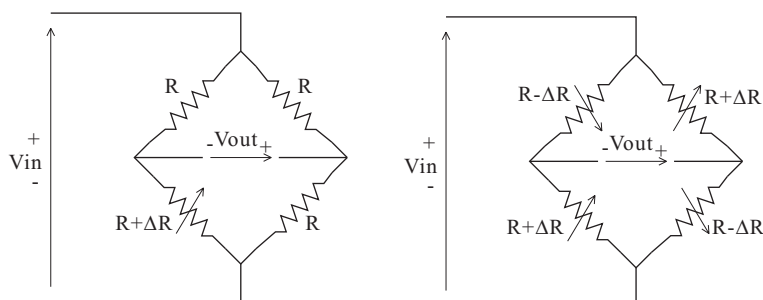


Figure 4.20: Resistors in a Wheatstone bridge with (left) one variable resistor, or (right) four variable resistors.

For resistors much longer than wide it is possible to write the relative change of resistance as :

$$\frac{\Delta R}{R} = \pi_l \sigma_l + \pi_t \sigma_t$$

where  $\pi_i$  is the piezoresistive coefficient and  $\sigma_i$  the stress component respectively, along the direction parallel to the current flow (l longitudinal) or perpendicular to it (t transverse). However the anisotropy in silicon, and actually in most crystals, makes it difficult to obtain the piezoresistive coefficients. Actually, all the physical parameters of silicon, like Young's modulus or conductivity, depends on the direction with respect to the crystal axes in which they are measured. Thus, a complete treatment of piezoresistivity will involve mathematical objects called tensors. Moreover, the piezoresistive coefficients in silicon depend on the type (n- or p-type) and concentration of doping impurities, being generally larger for p-type resistors, and finally, it depends also on temperature. However for the most important cases the expression can be found in the literature and for example for p-type resistors placed in a n-type substrate along the (110) direction, that is, parallel to the wafer flat in (100) wafers, we have  $\pi_l \approx 71.8 \cdot 10^{-11} \text{ Pa}^{-1}$  and  $\pi_t \approx -66.3 \cdot 10^{-11} \text{ Pa}^{-1}$ .

On a square membrane, for symmetry reasons, the stress in the middle of a side is essentially perpendicular to that side. Piezoresistor placed parallel or

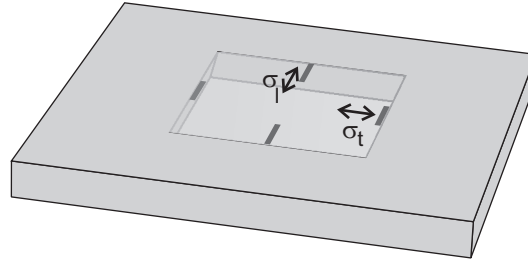


Figure 4.21: Typical position of piezoresistors for a square membrane on (100) Si wafer.

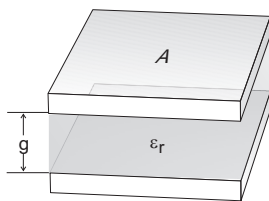
perpendicular to the side at that point will be, respectively, under transverse or longitudinal stress. If the membrane sides have been aligned with the (110) direction, the  $\pi_l$  and  $\pi_t$  are about the same magnitude but of opposite sign and the resistance of the two resistors under longitudinal stress in Figure 4.21 will increase when the membrane deforms while the resistance of the two resistors under transverse stress will decrease. It is thus possible to connect the four identical resistors in a full bridge configuration, as in Figure 4.20, and the bridge sensitivity simplifies to:

$$V_{\text{out}} \approx 70 \cdot 10^{-11} V_{\text{in}} \sigma$$

where  $V_{\text{in}}$  is the bridge polarization voltage and  $\sigma = \sigma_l = \sigma_t$  the maximum stress in the membrane. Although it seems really advantageous this configuration has seldom been used in practical devices because it suffers from a low manufacturability as the positioning of the resistors requires a very good accuracy. In general only one or two sensing resistors are used allowing simpler bridge balancing with trimmed external thick-film resistors.

Piezoresitivity is not only used for pressure sensor but find also application in acceleration or force sensors. Unfortunately, the simplicity of the method is counterbalanced by a strong dependence on temperature that has to be compensated for most commercial products by more complex circuitry than the elementary Wheatstone bridge, as described in section 5.3.3.

#### 4.4.2 Capacitive sensing



Capacitive sensing is a versatile sensing technique independent of the material used and it relies on the variation of capacitance appearing when the geometry of a capacitor is changing. Capacitance is proportional to

$$C \propto \epsilon_0 \epsilon_r \frac{A}{g}$$

where  $A$  is the area of the electrodes,  $g$  the distance between them and  $\epsilon_r$  the permittivity of the material separating them (actually, for a plane capacitor as

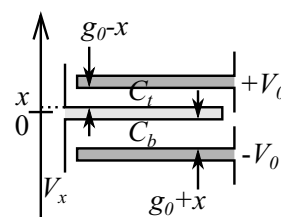
shown above, the proportionality factor is about 1). A change in any of these parameters will be measured as a change of capacitance and variation of each of the three variables has been used in MEMS sensing.

For example, accelerometers have been based on a change in  $g$  or in  $A$ , whereas chemical or humidity sensor may be based on a change of  $\epsilon_r$ .

If the dielectric in the capacitor is air, capacitive sensing is essentially independent of temperature but contrary to piezoresistivity, capacitive sensing requires complex readout electronics. Still the sensitivity of the method can be very large and, for example, Analog Devices used for his range of accelerometer a comb capacitor having a suspended electrode with varying gap. Measurement showed that the integrated electronics circuit could resolve a change of the gap distance of only 20 pm, a mere 1/5th of the silicon inter-atomic distance.

The conversion from capacitance to voltage can be obtained by using a bridge configuration but using AC voltage excitation. Then, instead of resistance as in the Wheatstone bridge we would then consider complex impedance of capacitor ( $Z_C = jC\omega$ ) and do the same math. A remaining issue would be to detect the amplitude of the signal (e.g., with a diode) but that principle would be sufficient for many applications. A more evolved principle has been used in some capacitive accelerometer from Analog Devices that highlights the interest of differential sensing and will be described in more details.

The designer have chosen to use the variation of the gap  $g$  between the electrodes as it may provide large sensitivity if the initial gap  $g_0$  is small enough. However, because the electrodes thickness was limited as they used surface micromachining, the capacitance was small and they connected many such capacitors in parallel using a fin-like structure with interpenetrated fingers for the mobile (rotor) and fixed (stator) parts.



If we zoom on a single rotor finger, we see that it is surrounded by two stator fingers that are polarized with AC voltage  $V_0$  and each constituting the fixed electrode of a variable capacitor where the rotor finger acts as the second movable electrode. We notice that when the rotor moves (e.g. for an accelerometer, because it is subjected to acceleration) one capacitance increases (decreasing gap) while the other decrease (increasing gap). Individually each of this capacitance varies very non-linearly with the position  $x$  and we have, for example for the top capacitor :

$$C_t = \epsilon_0 \epsilon_r \frac{A}{g_0 - x} = \epsilon_0 \epsilon_r \frac{A}{g_0} \frac{1}{1 - x/g_0} = C_0 \frac{1}{1 - x/g_0}$$

this expression can be linearized if we have  $x/g_0 \ll 1$ , and we get :

$$C_t \approx C_0 \left( 1 + \frac{x}{g_0} \right)$$

---

**Example 4.5** The resonant gate transistor.
 

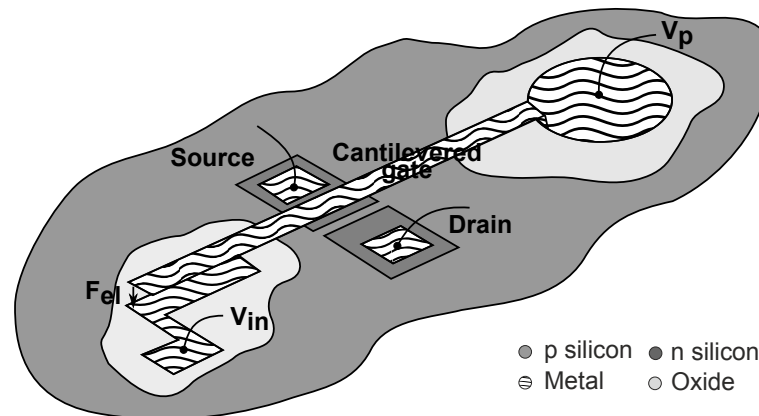
---

THE FIRST known MEMS, the Resonant Gate Transistor (RGT), has been developed by H. Nathanson in 1967[6] and was based on a variation of capacitive sensing. Actually the device was well ahead of his time, and used capacitive sensing amplification and capacitive actuation to yield a complete narrow bandpass filter.

Actually the input signal ( $V_{in}$ ) was setting the suspended beam into vibration with an amplitude depending directly on how close the excitation frequency of the signal was from the resonance frequency of the cantilever. At resonance, the vibrating cantilever polarized by  $V_P$  will come very close to the substrate, and act as the gate of a n-channel enhancement mode field effect transistor diffused into the p-type substrate. Actually the close presence of the positively charged gate would attract electrons from the silicon bulk into the gap between the source and drain electrodes, allowing the passage of a current that could be sensed by the electronic circuit.

If the gate remained far away because the input signal was not close to the beam resonant frequency, the FET transistor would remain blocked and no current would flow through.

In this way the device acted as a bandpass filter with a Q factor only determined by the mechanical properties of the beam and the viscous damping from air which could be very low if the system is placed in vacuum. Unfortunately this clever design was not used as the apparition of the operational amplifier made the issues encountered with inductor integration almost disappear, and filters were then build using only capacitors. Interestingly, high frequencies makes it harder and RGT like devices are studied again for replacing bandpass filters inside handphomes.



Let now consider the difference between the top and the bottom capacitance :

$$\begin{aligned}\Delta C &= C_t - C_b = \epsilon_0 \epsilon_r \frac{A}{g_0 - x} - \epsilon_0 \epsilon_r \frac{A}{g_0 + x} = \epsilon_0 \epsilon_r A \frac{2x}{g_0^2 - x^2} \\ &= 2C_0 \frac{x/g_0}{1 - (x/g_0)^2}\end{aligned}$$

which can be linearized if  $(x/g_0)^2 \ll 1$  as :

$$\Delta C \approx 2C_0 \frac{x}{g_0} \left( 1 + \left( \frac{x}{g_0} \right)^2 \right) \approx 2C_0 \frac{x}{g_0}$$

The interest ? Well, first the sensitivity has increased by a factor of 2 between the two cases, and besides the approximation is much less stringent, meaning that the non-linearity will be decreased. To convince yourself of this last point, imagine that the displacement  $x$  is 20% of the initial distance  $g_0$  – it is not sure that this case still qualify for  $x/g_0 = 0.2 \ll 1$ ... but clearly it would respect the  $(x/g_0)^2 = 0.04 \ll 1$  condition. Actually what we have done here is that we have removed the first order error in the approximation and now the error between the approximate (linear) formula and the exact formula is only of the second order.

However the last equation still depends on an approximation to be linear and on many geometrical parameters (buried in  $C_0$  expression) that could make it harder to go for industrialization. To overcome this issue the engineers came up with a more complex demodulation scheme (Figure 4.22) than the simple bridge configuration, that is based on using two AC excitation voltages of opposite sign (frequency 1 MHz) and performing the required amplitude detection with a multiplier.

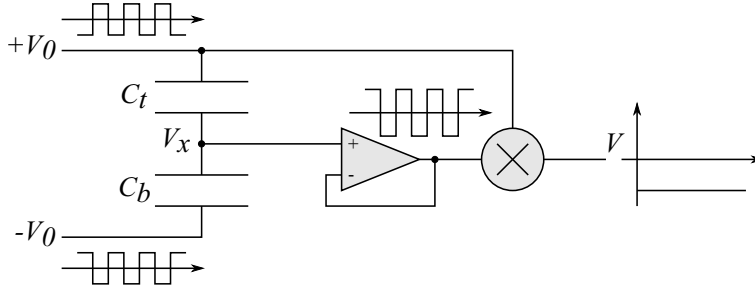


Figure 4.22: Demodulation principle for differential capacitance sensing.

The voltage  $V_x$  of the rotor electrode is floating and its value is found by considering the node there. Because of the buffer amplifier, no current goes out of this node and thus the total charge on the connected capacitors has to be 0, that is,  $Q_t + Q_b = 0$ , where  $t$  and  $b$  subscripts are used for the top and bottom capacitors as previously. Thus, using the capacitor relationship  $Q = CV$ , we can write

$$Q_t + Q_b = C_t(+V_0 - V_x) + C_b(-V_0 - V_x) = 0$$

getting  $V_x$  as :

$$V_x = V_0 \frac{C_t - C_b}{C_t + C_b}$$

Using the previous expression for the capacitance the voltage  $V_x$  is given by :

$$V_x = V_0 \frac{2C_0 \frac{x/g_0}{1-(x/g_0)^2}}{2C_0 \frac{1}{1-(x/g_0)^2}} = V_0 \frac{x}{g_0}$$

Although there is still dependence on one geometrical parameter  $g_0$  the  $C_0$  dependence is gone, and, more importantly, the expression is now fully linear and we may use displacement  $x$  as large as we want without any approximation ! The trick here has been to normalize the value of the difference by the average value of the capacitance, removing the factor in front of the expression. It should be noted that it is a general property and such principle could be used whenever a non-linear element response needs to be linearized.

The remaining issue is that the signal amplitude still need to be retrieved, which is performed by the multiplier after the buffer amplifier. Actually, by multiplying the signal  $V_x$  with  $GV_0$  ( $G$  is the gain of the multiplier) we actually get

$$V = GV_0^2 \frac{x}{g_0},$$

thus the signal becomes continuous ( $V_0$  is a symmetric bipolar signal of amplitude  $V_a$  thus  $V_0^2$  is simply a continuous signal of amplitude  $V_a^2$ ) with an amplitude directly proportional to the displacement.

### 4.4.3 Other sensing mechanism

A third commonly used transduction mechanism is based on piezoelectricity. The direct piezoelectric effect occurs when stress applied on a material induces the apparition of charge on its surface. Then a simple charge-to-voltage converter using an operational amplifier mounted as a transimpedance amplifier will convert this signal in a voltage. Silicon does not present piezoelectricity but crystalline quartz has a large piezoelectric coefficient and other material like ZnO or PZT can be deposited in thin films possessing piezoelectric properties. The advantage of the piezoelectric effect is that it is reversible and can be used to sense stress but also for actuation. Actually a difference of potential applied on two sides of a piezoelectric layer will induce its deformation. Thus piezoelectric material can be excited in vibration and the vibration sensed with the same structure. This has been the heart of the quartz watch since its invention in the 1970's, but it is also used for different inertial MEMS sensor like gyroscope.

Magnetic sensing, although less often used, has its supporters mainly because it is a non-contact sensing mechanism with a fairly long range. Its main application has to be found in the (giant)magneto-resistive effect used inside the hard-disk



head. However other uses of magnetic sensing have been tested and for example some sensors have been based on the Hall effect taking advantage of the simplicity to manufacture this sensing element.

## 4.5 Actuator technology

Since the industrial revolution we have understood that machines can perform tasks with more force and endurance than humans. Bulldozers moving around with their huge engine and pushing big rocks with their powerful pneumatic actuators are probably a good example of what big machines can do. But what will be the function of a micro-sized actuator?

Type		Force	Stroke	Efficiency	Manuf.
Electromagnetic		+	+	-	-
Electrostatic	Gap-closing	0	-	+	+
	Comb-drive	-	0	+	+
	SDA	0	+	+	0
Piezoelectric		+	-	+	-
Thermal	Bimorph	+	+	0	0
	Heatuator	0	0	0	+
	Shape memory alloy (SMA)	+	+	+	-
	Thermo-fluidic	+	-	0	0

Table 4.13: Comparison of common micro-mechanical actuators.

The main parameters useful to describe an actuator are its force and its stroke. However we have seen previously that all forces decrease with the scale, thus we can not expect to move big rocks around - but only micro-rocks. The micro-actuators are currently used to act on micro-object, typically one part of a MEMS device, and generate forces in the micro to milli Newton range with a stroke from a few  $\mu\text{m}$  to several hundreds  $\mu\text{m}$ . It would be interesting to have enough force and stroke to allow actuator to help interface human and machine by providing force feedback for example, but micro-actuators are still unable to do that properly. Still a wide range of principles exists that would transform internal energy of a system (usually electrical energy) to energy in the environment (in the case of MEMS, generally mechanical energy). Sometimes the conversion from electric energy to mechanical energy is direct but often another intermediate energy form is used. For example, the heatuator, a form of thermal actuator, uses current to

generate heat which in turn becomes strain and displacement.

The MEMS actuators can be conveniently classified according to the origin of their main energy form. In Table 4.13 we compare the most common MEMS actuators, where Efficiency refers to the loss existing in the actuator conversion of electrical energy to mechanical energy and Manuf. is the manufacturability or the easiness of mass micro-fabrication.

### 4.5.1 Magnetic actuator

Electromagnetic actuation is recognized for providing most of the actuators used in house appliances, toys, watches, relays... The principle of electromagnetic motor is well known and it is tempting to miniaturize such a versatile device to use it in the micro-world. However an electromagnetic motor with its coils, armature and bearings prove a tremendous task for micro-fabrication and so far nobody has been able to batch produced a motor less than 1mm diameter.

Still magnetic actuation has many proponents and some version of magnetic linear actuator have been used in different devices. Such a mobile armature actuator is shown in Figure 4.23, where by increasing the current in the coil the mobile armature is attracted along the x direction to align with the fixed armature.

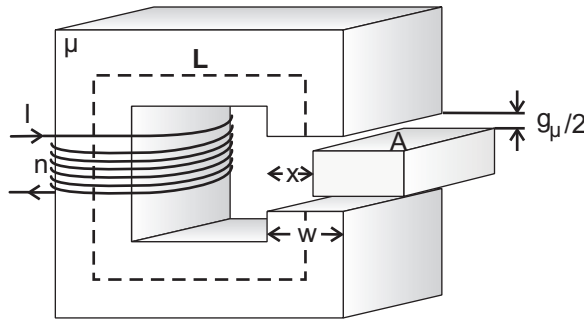


Figure 4.23: Mobile armature linear magnetic actuator.

The magnetic force produced on the mobile armature is linked to the change of reluctance and is given approximately by [14] :

$$F_{\text{ma}} = \frac{(nI)^2}{2w} \left( \frac{\mu_0 A}{g_\mu + \mu_0 L/\mu} \right)$$

From this equation it is clear that the force is non linear with the current, and assuming a constant resistance for the coil, the force will also depend on the square of the coil voltage.

Although this force does not scale very favorably, the possibility to increase the current at small scale, because the heat can be dissipated more quickly, still allows producing relatively strong force. However the main difficulty preventing the widespread use of this type of actuator in a MEMS component is the fabrication of the

coil. In that case the most convincing approach proposed so far are most probably those using a hybrid architecture, where the magnetic circuit is fabricated using micro-fabrication but the coil is obtained with more conventional techniques and later assembled with the MEMS part. Actually some design have shown that the coil does not need to be microfabricated at all and can be placed in the package, taking benefit of the long range action of the magnetic field.

Finally it should be noted that magnetic actuation can be used in conjunction with ferro-magnetic material to provide bistable actuator where two positions can be maintained without power consumption. A permanent magnet placed in the package is used to maintain the magnetized ferro-magnetic material in place. Then, when we send a current pulse of the right polarity in a coil wound around the ferro-magnetic material we invert its magnetization and the actuator switch to its second state. NTT has been producing since at least 1995 a fiber optic switch based on a moving fiber with a ferro-nickel sleeve that has two stable positions in front of two output fibers [15]. The device will consume power only during the brief time where the current pulse is sent and can maintain its position for years.

### 4.5.2 Electrostatic actuator

A physical principle that leads itself well to integration with MEMS technology is electrostatics. Actually by applying a potential difference between two electrodes, they develop charges of opposite sign and start attracting each other through the Coulomb's force. This principle has known several applications among which, the comb-drive actuator, the gap-closing actuator and the scratch drive actuator are the most commonly used (Figure 4.24). To derive an expression of the force developed by such actuator we will use the principle of virtual work. This principle states that for energy conserving systems (no dissipation) the potential energy of the system changes as the work of internal forces. This is another way of saying that a system tends towards a state of minimal energy – in mathematical terms it says something like:

$$\vec{F} = -\overrightarrow{\text{grad}}U$$

with  $\overrightarrow{\text{grad}} = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})^t$  a vector composed of the partial derivative along each axis. We note the minus sign in front of the gradient, saying that the direction of the force is opposite to the direction of increase (gradient is positive if the function rises) of energy: that is, the force is leading the system toward a minimum of stored energy.

In this case, as simple electrostatic systems are capacitors, the energy<sup>11</sup> stored in the system upon application of an external voltage is

$$U_C = \frac{1}{2}CV^2.$$

---

<sup>11</sup>Actually we have to use here the co-energy  $\int v idt = \int v dq = \int v C dv = 0.5CV^2$

Thus any variation of the internal (potential) energy of the capacitor due to some change in its geometry can be attributed to the (virtual) work of a force. Thus the force developed between the two electrodes becomes proportional to the gradient (spatial derivative) of the energy :

$$\vec{F}_{\text{elec}} \propto V^2 \vec{\text{grad}} C.$$

This fundamentally shows that electrostatic actuators develop force non-linear with voltage and proportional to the gradient of the capacitance. The  $V^2$  depen-

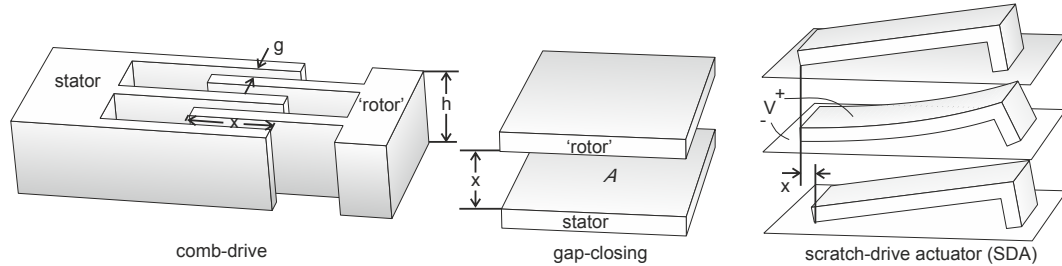


Figure 4.24: Different type of electrostatic actuators.

dence shows that potentially (pun intended!) the force can be quite large if we are able to increase the voltage substantially. However as we know – lightning strike during a storm would make a good recall – there is in air a maximum value of the electric field that would cause ionization and a destructive short circuit, preventing the voltage to be too high. Interestingly it has been shown at the end of the XIX<sup>th</sup> century by F. Paschen that the ionization field is not constant and actually depends on the air pressure and on the gap between the electrodes. Actually if one think at the ionization phenomena as an avalanche process (a single accelerated electron hits an atom releasing two such electrons which in turns get accelerated and hits two other atoms, liberating four electrons, etc) this dependence of the ionization field can be relatively simply understood: if the mean free path between collision is much larger than the electrodes gap then ionization may not occur. This can be obtained in two ways: either the mean free path is increased by reducing the pressure (that is, at constant temperature, the density of gas molecules), or alternatively the gap between the electrodes is reduced. At micro-scale it is this second phenomena that occurs and allow for much higher electric field than at large scale. This is the main contributor to the interest of electrostatic force at microscale, making it able to pack almost as much energy in a finite volume (energy density) than electromagnetic energy.

However avalanche effect is not the only contributor to conductivity and with high enough field and even in vacuum electron emission will occur through other mechanism (e.g., field emission at surface asperities) and finally result in arcing. Experimental investigation of this effect has lead to the elaboration of modified Paschen's curves (Figure 4.25) showing an estimate of the breakdown voltage as a function of the gap that take into consideration the Paschen's curve and the

vacuum emission. These critical curve will vary with the gas used, its pressure,

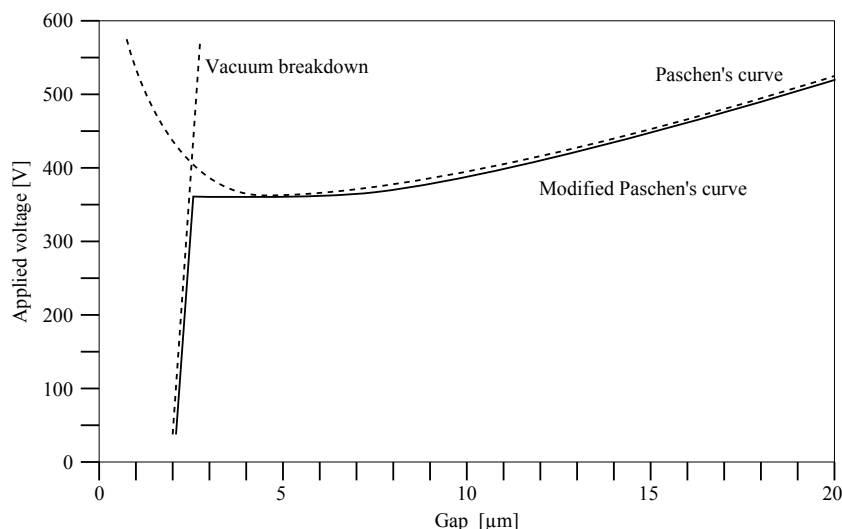


Figure 4.25: Maximum admissible voltage in electrostatic actuators following modified Paschen's law at 1 atm.

the electrodes material and the presence of asperities on the electrodes (including at its edge). Still, we see that for obtaining maximum force there is little incentive to get a gap below 2-3  $\mu\text{m}$ , and that a typical voltage of several 100 V may be used for micro-electrodes. In practice test will be required to ascertain the real critical voltage, but robust development will try to use gap of 4  $\mu\text{m}$  to avoid random effect due to field emission while keeping the maximum voltage below 200 V.

The comb-drive actuator (Figure 4.24-left) was invented by W. Tang [16] at UC Berkeley and it generally allows motion in the direction parallel to the finger length. The capacitance can be obtained by considering each side of a finger behaves as a parallel plate capacitor, giving for each finger a capacitance of  $C \approx 2\epsilon_0 hx/g$ . Taking the gradient, the force produced by  $n$  fingers in the rotor is approximately given by

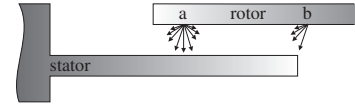
$$F_{\text{cd}} \approx n\epsilon_0 \frac{h}{g} V^2$$

where we see the expected dependence with the square of the voltage and notice that it is independent of the displacement  $x$ . The proportionality factor is  $\epsilon_0$ , a small quantity indeed, hinting to a small force generated per finger, in the order of a few 10 nN. Of course the number of fingers can reach 100 or more and the actuator aspect-ratio can be made larger (i.e., increase  $h/g$ ) to increase its force proportionally. This actuator has been used repeatedly in MEMS component, for example in the original Analog Devices accelerometer or in the fiber optic switch from Sercalo.

The origin of the force parallel to the electrodes results from the unbalance of the Coulomb's force along some part of the finger. Actually the rotor charges

located in (a) will experience a symmetrical attraction resulting in an absence of net force, whereas, charges located in (b) will, as a result of the unbalanced attraction, create a net force pulling the rotor parallel to the stator. We note that the Coulomb's force also results in a force perpendicular to the surface, however the motion toward the stator is prevented by the rotor suspension and by the balancing force of the stator finger placed on the other side of the rotor finger.

The gap-closing actuator (Figure 4.24 center) actually makes use of this force perpendicular to the electrodes surface and usually delivers a larger force (proportional to  $A$ ). Actually the capacitance is now expressed as  $C \approx \epsilon_0 A/x$ , resulting in a force again non linear with the applied voltage, but additionally depending on the displacement  $x$ :



$$F_{gc} \approx \epsilon_0 \frac{A}{2x^2} V^2.$$

As this force is unidirectional (changing the voltage sign does not change the force direction), a reversible actuator will need a restoring force to bring it back to its original position. This is usually obtained with a spring (usually a bending beam) that will be used to polarize and retain the rotor electrode as seen in Figure 4.26. To find the rest position, we write the force equilibrium,

$$F_{gc} + k(g - x) = 0$$

where  $k(g - x)$  is the magnitude of the upward directed spring force. Thus we get a third order equation relating the position  $x$  with the applied voltage  $V$ .

$$x^3 - gx + \epsilon_0 \frac{A}{2k} V^2 = 0$$

Using Cardan's method, the roots can be found and we get:

$$x_1 = \frac{2}{3}g \cos \left( \frac{1}{3} \arccos \left( 1 - \frac{2V^2}{V_{\text{pull-in}}^2} \right) \right) + \frac{g}{3} \quad (4.15)$$

$$x_2 = \frac{2}{3}g \cos \left( \frac{1}{3} \arccos \left( 1 - \frac{2V^2}{V_{\text{pull-in}}^2} \right) + \frac{4\pi}{3} \right) + \frac{g}{3} \quad (4.16)$$

plus another root yielding  $x < 0$  that is unphysical. The two physical roots of this equation have been plotted in Figure 4.26. We note that instead of solving the third order polynomial equation, we may plot  $V$  as a function of  $x$  for  $0 < x < g$ . The solution shows that, the rotor position can only be controlled over a limited range, and actually one root corresponds to a completely unstable position. Actually, when the voltage is increased the rotor slowly moved toward the stator electrode but as soon as the rotor has moved by one third of the original gap width ( $g$ ), snap-in suddenly occurs and the rotor comes into contact with the stator (in the figure

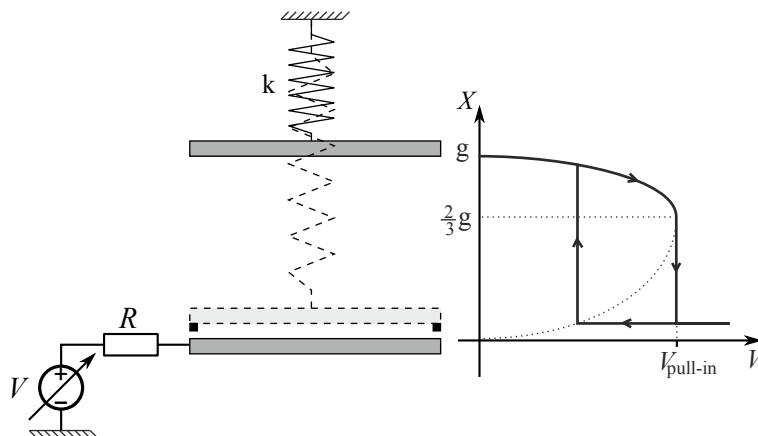


Figure 4.26: Voltage-Displacement curve for a gap-closing actuator showing the pull-in phenomena.

we show two blocks in black used to prevent the contact between the electrodes and a short-circuit, the position of these blocks will determine the pull-out voltage when the voltage is decreased). The pull-in voltage is given by:

$$V_{\text{pull-in}} = \sqrt{\frac{8}{27} \frac{kg^3}{\epsilon_0 A}}$$

where  $g$  is the original gap width and  $k$  the rotor suspension spring constant. This behavior can be advantageous if the actuator is used for bi-stable operation, but as here, preventive measures should be taken to avoid electrodes short-circuit. Actually, the actuator behind the Texas Instruments' DLP is a gap-closing electrostatic actuator working in torsion with the two stable states position fixed by resting posts. By biasing the actuator at a voltage in the middle of the hysteresis curve, it needs only a small swing of voltage to allow a robust bi-stable actuation.

The scratch drive actuator (Figure 4.24 right) has been invented by T. Akiyama [17] and although it is actuated by a varying electrostatic field, the friction force is the real driving force. As we can see in the diagram, as the voltage is applied, the electrostatic energy is stored in the SDA strain while its front part, the bushing, bends. When the voltage is released, the strain energy tends to decrease and the elasticity of the bushing returns it to its rest orientation producing displacement. The main advantage of this actuator is that it is able to produce a rather large force (100  $\mu\text{N}$ ), which can be even increased by connecting multiple actuators together. Actually the SDA has been used as an actuator in the 2D optical switch matrix that was developed by Optical Micro Machines (OMM) and which received the stringent Telcordia certification.

Electrostatics can also be used to move liquids. It is based on two phenomena: electro-hydrodynamic which works with non-conductive fluids and electro-osmosis which works with conductive fluids. Electro-osmosis pumps are of larger significance because biological fluids are actually solute with different ions (salts) and

are thus conductive.

In an electro-osmosis pump a stationary electric field is applied along a channel and result in an overall motion of the conductive fluid. Although the liquid is globally neutral, this motion can happen because of a so-called double layer of charged particles near the walls of the channel. In general with conducting fluids

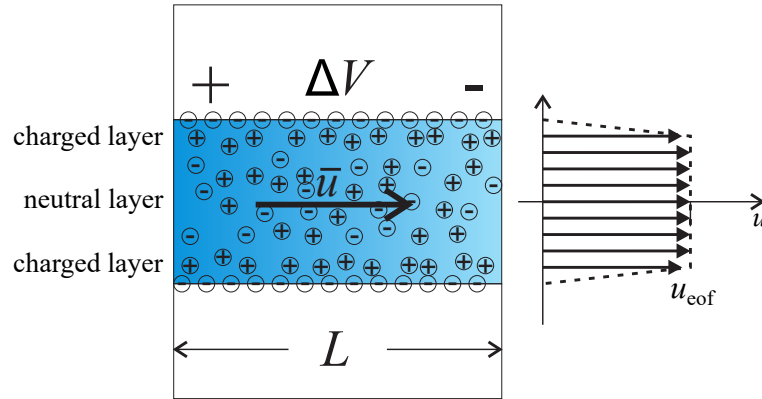


Figure 4.27: Electro-osmosis flow in a channel.

a surface charge appears on the insulating walls, usually made of polymer or glass, resulting in a negative potential (the zeta potential  $\zeta$ ) ranging generally between -20 mV and -200 mV. Those trapped charge attracts positive ions in the solution close to the wall creating the double layer of a few nm thick, as shown in Figure 4.27. Then, the external electrostatic field makes this layer move, entraining all the liquid in the channel, because of the shear force present in liquids with non-zero viscosity. The flow velocity profile, contrary to pressure driven flow, is uniform across the channel (called a “plug flow”) and we have:

$$\bar{u}_{\text{eof}} = u_{\text{eof}} = \frac{\epsilon \zeta}{\eta} E \quad (4.17)$$

where  $\epsilon$  is the dielectric constant of the fluid and  $\eta$  its viscosity. Actually the proportionality constant between the velocity and the field is usually rather small ( $\mu_{\text{eo}} = \epsilon_0 \epsilon_r \zeta / \eta$ ,  $\mu_{\text{eo}}$  electro-osmotic mobility) and in practice the voltage used for electro-osmosis flow across a 10 mm-long channel will be in the order of 1000 V.

### 4.5.3 Piezoelectric actuator

At the end of the 19<sup>th</sup> century Pierre and Jacques Curie discovered that certain materials produce electrical charges at their surface when they are submitted to an external force – the so called direct piezoelectric effect. Reciprocally, when these materials are subjected to an electric field they contract or expand – the so called converse piezoelectric effect. The materials themselves are called piezoelectric materials and are natural transducers between electrical and mechanical domains. They have thus been used inside different mechanical sensors and actuators and are



particularly interesting for micro-scale actuation because they have a high power density, producing rather large force for small volume. However, the deformation being induced in the rigid bulk of the material, the magnitude of the deformation remains small requiring clever designs to obtain actuators with large stroke. Fundamentally the origin of piezoelectricity is linked with the absence of center

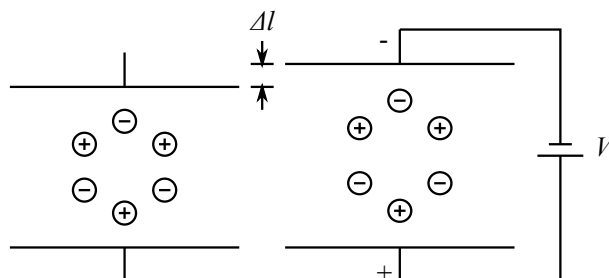


Figure 4.28: Converse piezoelectric effect in a non-centrosymmetric crystal.

of inversion symmetry in the ionic crystal – at the nano-scale we have a dipole (polarization) that for the converse effect will be influenced by an external field and because of the regular charge repartition in the crystal won't cancel, inducing a macroscopic deformation.

Piezoelectricity shows coupling between electrical and mechanical variables, and require both domain for a complete description. As we have seen previously the elasticity of a crystal can be described by relating the strain  $\epsilon$  to the stress<sup>12</sup>  $\sigma$  using the compliance tensor  $\epsilon = S\sigma$ . Moreover, we know that the electric displacement for a linear polarization proportional to the electric field is given by  $\vec{D} = \epsilon\vec{E}$ , where, for anisotropic materials, the permittivity  $\epsilon$  is a second rank tensor (a  $3 \times 3$  matrix).

The piezoelectric effect is linear and for small deformation can be considered to be independent of the compliance and the permittivity of the material. Thus we can write the combination of piezoelectricity and elasticity as:

$$\epsilon = S^E\sigma + d^t\vec{E} \quad (4.18)$$

$$\vec{D} = d\sigma + \epsilon^\sigma\vec{E} \quad (4.19)$$

with  $S^E$  the compliance at  $\vec{E} = 0$ ,  $d^t$  the transpose of the piezoelectric charge tensor  $d$ ,  $\vec{E}$  the electric field,  $\vec{D}$  the electric displacement and  $\epsilon^\sigma$  the permittivity tensor at  $\sigma = 0$ .

For the converse piezoelectric effect, the transpose of the piezoelectric charge tensor  $d^t$  relates the strain (second rank tensor) to the electric field (a vector or first rank tensor) and is thus a third rank tensor, which should require  $3^3 = 27$  terms. However, the existing symmetry of the tensor with respect to stress and

<sup>12</sup>As described previously these are both second rank tensors, but written as a 6 components vector by considering symmetry and using index contraction.

the index contraction rule allows for the  $d^t$  tensor to be simply represented as a  $6 \times 3$  matrix as:

$$d^t = \begin{bmatrix} d_{11} & d_{21} & d_{31} \\ d_{12} & d_{22} & d_{32} \\ d_{13} & d_{23} & d_{33} \\ d_{14} & d_{24} & d_{34} \\ d_{15} & d_{25} & d_{35} \\ d_{16} & d_{26} & d_{36} \end{bmatrix}$$

Again, symmetries results in many piezoelectric coefficient  $d_{ij}$  to be zero in most crystals when we use the (X=A,Y=B,Z=C) crystallographic system of coordinate. In this case, for quartz (trigonal crystal) we have:

$$d^t = \begin{bmatrix} d_1 & 0 & 0 \\ -d_1 & 0 & 0 \\ 0 & 0 & 0 \\ d_2 & 0 & 0 \\ 0 & -d_2 & 0 \\ 0 & -2d_1 & 0 \end{bmatrix}$$

with  $d_1 = d_{11} = -2.3 \cdot 10^{-12}$  C/N and  $d_2 = d_{14} = -0.67 \cdot 10^{-12}$  C/N. For zinc oxide (ZnO), a material from the hexagonal class different from quartz that can be deposited by sputtering, we get in the (A,B,C) system of coordinate:

$$d^t = \begin{bmatrix} 0 & 0 & d_1 \\ 0 & 0 & d_1 \\ 0 & 0 & d_2 \\ 0 & d_3 & 0 \\ d_3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

with  $d_1 = d_{11} = -5.4 \cdot 10^{-12}$  C/N,  $d_2 = d_{33} = 11.6 \cdot 10^{-12}$  C/N and  $d_3 = d_{42} = -11.3 \cdot 10^{-12}$  C/N.

In practice the most commonly used materials are artificial ceramic (e.g., Lead Zirconate Titanate or PZT) which, although they present at nano-scale the right lack of center of symmetry, are not naturally piezoelectric because they are polycrystalline. Actually, in each grain the random orientation of the polarization

cancels each other resulting in a lack of observable piezoelectric effect. For circumventing this problem, the materials is heated (close to the Curie temperature where reorganization can easily occur) and submitted to a large electrical field, an operation called poling, that orients the polarization of all the nano-crystallite in the same direction. When the material is cooled down, most of this orientation is preserved and the resulting material shows macroscopic piezoelectric effect. The PZT crystal symmetry is the same than ZnO but the piezoelectric coefficient are much larger making this crystal very interesting for lower voltage operation. We have for poled PZT  $d_1 = d_{11} = -123 \cdot 10^{-12}$  C/N,  $d_2 = d_{33} = 289 \cdot 10^{-12}$  C/N and  $d_3 = d_{42} = 496 \cdot 10^{-12}$  C/N.

Besides the PZT it is possible to deposit by sputtering thin-film crystalline material like AlN and ZnO that present interesting piezoelectric properties without poling. Alternatively it is also possible in certain cases to work on piezoelectric substrate like quartz, gallium arsenide (AsGa) or lithium niobate (LiNbO<sub>3</sub>). Most of the time, the purpose of the actuator is to generate acoustic waves (vibration) into the materials but they are rarely used to produce mechanical work.

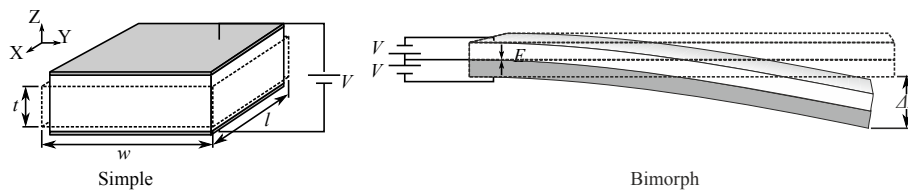


Figure 4.29: Simple and bimorph piezoelectric actuators.

For the simple actuator in Figure 4.29, if we apply voltage along the crystal C-axis in a ZnO layer, the field along C is  $V/t$  (and 0 along the other axes) and, placing ourselves in the  $(X=A, Y=B, Z=C)$  system of coordinate, we obtain a change in dimension according to (4.18) as:

$$\begin{bmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \\ \gamma_{XY} \\ \gamma_{YZ} \\ \gamma_{ZX} \end{bmatrix} = \begin{bmatrix} 0 & 0 & -5.4 \\ 0 & 0 & -5.4 \\ 0 & 0 & 11.6 \\ 0 & -11.3 & 0 \\ -11.3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} 10^{-12} \cdot \begin{bmatrix} 0 \\ 0 \\ V/t \end{bmatrix}$$

Thus

$$\begin{bmatrix} \epsilon_X \\ \epsilon_Y \\ \epsilon_Z \\ \gamma_{XY} \\ \gamma_{YZ} \\ \gamma_{ZX} \end{bmatrix} = \begin{bmatrix} -5.4 \cdot 10^{-12} \frac{V}{t} \\ -5.4 \cdot 10^{-12} \frac{V}{t} \\ 11.6 \cdot 10^{-12} \frac{V}{t} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Finally, using the expression of longitudinal strain ( $\epsilon = \Delta L/L$ ) we get  $\Delta w = -5.4 \cdot 10^{-12} V \frac{w}{t}$ ,  $\Delta l = -5.4 \cdot 10^{-12} V \frac{l}{t}$  and  $\Delta t = 11.6 \cdot 10^{-12} V$ . We notice that the thickness change  $\Delta t$  is independent of the layer thickness  $t$  – for a voltage of 100 V it will be about 1.1 nm.

It is clear that an actuator simply based on longitudinal expansion will have a very short stroke, and normally piezoelectric actuator rely on stacking many layers of materials and electrodes or on bimorph-type structure. In this last structure shown in Figure 4.29, because the two layers are polarized with fields of opposite signs, one will expand and the other contract along the Y-axis<sup>13</sup> creating large stress at the interface and resulting in bending of the beam with a large stroke  $\Delta x$ .

#### 4.5.4 Thermal actuator

The thermal energy used by this class of MEMS actuator comes almost invariably from the Joule effect when a current flows through a resistive element. These actuators are generally relatively strong and their main drawback is most probably their speed, although at micro-scale the heat is quickly radiated away and operating frequency up to 1 kHz can be achieved.

Bimorph actuators are the most common type of thermal actuator. The bi-material actuator, well known from the bimetallic version used in cheap temperature controller, and the heatuator (Figure 4.30) are both bending actuator where bending is induced by a difference of strain in two members connected together.

The bi-material actuator obtains this effect by using two different materials with different coefficients of thermal expansion that are placed at the same temperature resulting in a misfit strain  $\epsilon_m = (\alpha_1 - \alpha_2)\Delta T$ . The curvature of a bi-material actuator is given by:

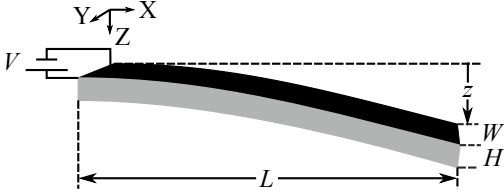
$$\kappa = \frac{6E_1E_2(h_1 + h_2)h_1h_2\epsilon_m}{E_1^2h_1^4 + 4E_1E_2h_1^3h_2 + 6E_1E_2h_1^2h_2^2 + 4E_1E_2h_2^3h_1 + E_2^2h_2^4} \quad (4.20)$$

where  $E_i$  is the Young's modulus for the two materials and  $h_i$  their thickness. Actually the difference in thermal expansion existing for materials deposited at

<sup>13</sup>It will also happen along the X-axis but with much less bending as the beam is narrow.

**Example 4.6** Dynamic model of a bimorph piezoelectric actuator

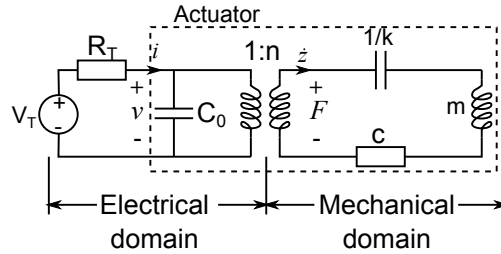
WE WANT to analyze a piezoelectric actuator obtained by depositing on a silicon cantilever a ZnO layer of thickness  $t$  with its C axis perpendicular to the surface (that is along the Z axis). We will be using a model similar to what is described by Senturia[14].



A voltage applied on the electrodes will induce stress at the ZnO/Si interface through the converse piezoelectric effect, creating a bending moment, effectively bending the cantilever and deflecting the tip. At the same time the direct piezoelectric effect will change the polarization inside the piezoelectric layer because of its

deformation, modifying the capacitance of the actuator.

The piezoelectric actuator is linear and thus the dynamic of this actuator can be modeled with the formalism developed in Section 2.5 – actually we are trying to calculate the values  $C_0$ ,  $n$ ,  $k$ ,  $m$ ,  $c$  of the circuit elements appearing in Example 2.10. As we have seen, applying the electrical field  $E_Z$  along the Z-axis (the C axis of the crystal) creates expansion along the Z direction. As the film is free to extend in this direction  $\sigma_Z = 0$ . However the contraction occurring along the X and Y directions are prevented because the film is placed on a rigid substrate ( $\epsilon_X = \epsilon_Y = 0$ ). Thus we have:



$$\begin{bmatrix} 0 \\ 0 \\ \epsilon_Z \\ \gamma_{XY} \\ \gamma_{YZ} \\ \gamma_{ZX} \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & 0 & 0 & 0 \\ s_2 & s_1 & s_3 & 0 & 0 & 0 \\ s_3 & s_3 & s_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & s_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(s_1 - s_2) \end{bmatrix} \cdot \begin{bmatrix} \sigma_X \\ \sigma_Y \\ 0 \\ \tau_{XY} \\ \tau_{YZ} \\ \tau_{ZX} \end{bmatrix} + \begin{bmatrix} 0 & 0 & d_1 \\ 0 & 0 & d_1 \\ 0 & 0 & d_2 \\ 0 & d_3 & 0 \\ d_3 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ V/t \end{bmatrix}$$

with  $s_1 = 7.79 \cdot 10^{-12} \text{Pa}^{-1}$ ,  $s_2 = -3.63 \cdot 10^{-12} \text{Pa}^{-1}$ ,  $s_3 = -2.12 \cdot 10^{-12} \text{Pa}^{-1}$ ,  $s_4 = 6.28 \cdot 10^{-12} \text{Pa}^{-1}$  and  $s_5 = 24.7 \cdot 10^{-12} \text{Pa}^{-1}$ . We use the two equations along X and Y and write

$$\begin{cases} 0 = s_1 \sigma_X + s_2 \sigma_Y + d_1 \frac{V}{t} \\ 0 = s_2 \sigma_X + s_1 \sigma_Y + d_1 \frac{V}{t} \end{cases}$$

---

**Example 4.6** Dynamic model of a bimorph piezoelectric actuator (continued)

---

and obtain  $\sigma_X = \sigma_Y = -\frac{d_1}{s_1+s_2} \frac{V}{t}$  showing that we have a bi-axial stress in the cantilever.

Because the piezoelectric thin film is much thinner than the silicon beam ( $H \gg t$ ), we can assume the neutral axis is still in the middle of the beam and the stress is just at its surface. We also understand that the bending of the cantilever will mostly be along its length and ignore the bending along Y. Then, neglecting the stiffness of the piezoelectric layer, the bending moment  $M$  is simply given by considering the force  $\sigma_X tW$  at a distance  $H/2$  of the neutral axis:

$$M \approx \sigma_X tW \frac{H}{2} = -\frac{d_1}{s_1+s_2} \frac{HW}{2} V$$

it is again independent of  $t$ , the thickness of the piezoelectric layer. The moment is also independent of  $x$  thus the beam is under pure bending moment along Z. Using Eq. (4.1) and  $I = H^3W/12$  we write:

$$\frac{d^2z}{dx^2} = \frac{s_1 M}{I} = -\frac{s_1 d_1}{s_1+s_2} \frac{6}{H^2} V$$

Thus, the deflection of the actuator  $z(x)$  is obtained by simply integrating twice the previous expression and assuming that  $z(0) = 0$  and the slope  $dz/dx$  is also 0 at  $x = 0$ :

$$z(x) = \frac{d_1 s_1}{s_1+s_2} \frac{3}{H^2} V x^2$$

The beam assumes a parabolic profile, and the deflection at the tip  $x = L$  is:

$$z_L = \frac{d_1 s_1}{s_1+s_2} \frac{3}{H^2} L^2 V$$

Noticing that in the mechanical domain considering the spring in quasi-static we have  $F = kz_L$  and that the transformer implies that  $F = nV$  we can write that:

$$nV = kz_L \Rightarrow z_L = \frac{n}{k} V$$

By identifying the two expressions obtained above for  $z_L$ , we get:

$$\frac{n}{k} = \frac{d_1 s_1}{s_1+s_2} \frac{3}{H^2} L^2$$

We turn now our attention to the electric domain equations, assuming that the cantilever deformation does not change significantly the stress in the piezoelectric layer. Thus we can write the equation for the direct piezoelectric effect in this layer as:

---

---

**Example 4.6** Dynamic model of a bimorph piezoelectric actuator (continued)

---

$$\begin{bmatrix} D_X \\ D_Y \\ D_Z \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_3 & 0 & 0 \\ d_1 & d_1 & d_2 & 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} \sigma_X \\ \sigma_Y \\ 0 \\ \tau_{XY} \\ \tau_{YZ} \\ \tau_{ZX} \end{bmatrix} + \begin{bmatrix} \varepsilon_X & 0 & 0 \\ 0 & \varepsilon_Y & 0 \\ 0 & 0 & \varepsilon_Z \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ V/t \end{bmatrix}$$

The Gauss' law expressed with the electric field displacement  $\vec{D}$  states that  $\oint \vec{D} \cdot \vec{n} dA = Q$  the charge in the system. The electrodes are in (X,Y) parallel planes, thus the normal to the electrodes is along Z and the Gauss' integral simplifies as  $D_Z WL = Q$ . The two other components of the electric displacement ( $D_X$  and  $D_Y$ ) do not contribute to the electrodes polarization. From the direct piezoelectric effect equation, we find that  $D_Z = d_1 \sigma_X + d_1 \sigma_Y + \varepsilon_Z \frac{V}{t}$ , thus using the previous expression for the biaxial stress in the piezoelectric layer we obtain:

$$Q = D_Z WL = \left( 2 \frac{d_1^2}{c_1 + c_2} \frac{WL}{t} + \varepsilon_Z \frac{WL}{t} \right) V$$

There are two contributions to the capacitance of the electrodes: a first capacitance term due to the charge appearing through the direct piezoelectric effect and the biaxial stress in the piezoelectric layer, and a second capacitance term arising from the charge displacement in the dielectric. This second term  $\varepsilon_Z \frac{WL}{t}$  is the capacitance from the electric domain and is the  $C_0$  of the model. The first term  $2 \frac{d_1^2}{c_1 + c_2} \frac{WL}{t}$  is actually the capacitance of the mechanical domain  $1/k$  seen from the electrical side. As an impedance connected on a transformer secondary is divided by  $n^2$  when it is seen from the primary, the equality between these two capacitances is written as:

$$\frac{n^2}{k} = 2 \frac{d_1^2}{s_1 + s_2} \frac{WL}{t}$$

Thus taking the ration between this expression, and the expression found previously in the mechanical domain for  $n/k$ , we get an expression for  $n$  and  $k$ :

$$n = \frac{2d_1}{3s_1} \frac{WH^2}{tL} \quad k = \frac{2}{3} \frac{s_1 + s_2}{s_1^2} \frac{WH^2}{tL}$$

Now, we turn our attention to the value of the effective mass  $m$  appearing in the model. We will estimate it by using Rayleigh's method. The cantilever bends under a parabolic profile  $z = ax^2$ , where  $a = \frac{d_1 s_1}{s_1 + s_2} \frac{3}{H^2} V$ .

---

**Example 4.6** Dynamic model of a bimorph piezoelectric actuator (end)

We use this shape to approximate the vibrating mode shape under varying stress as

$$z(x, t) = ax^2 \sin(\omega t)$$

The velocity along the beam is then given by:

$$\dot{z}(x, t) = ax^2 \omega \sin(\omega t)$$

Thus the kinetic energy stored in the beam is:

$$\begin{aligned} K_b &= \frac{1}{2} \int_V \rho \dot{z}^2 dV = \frac{\rho}{2} \int_A dA \int_0^L \dot{z}^2 dx = \frac{1}{2} \rho HW \int_0^L a^2 x^4 \omega^2 \sin^2(\omega t) dx \\ &= \frac{1}{10} \rho HW a^2 L^5 \omega^2 \sin^2(\omega t) \end{aligned}$$

The bending moment in the beam is given by  $M(x, t) = EI \frac{d^2 z}{dx^2} = 2 \frac{I}{s_1} a \sin(\omega t)$  where the Young's modulus along  $Z$  is taken as  $E = 1/s_1$ . We evaluate the elastic energy as shown in Section 4.3.2 for a beam in pure bending:

$$\begin{aligned} U_b &= \frac{1}{2} \int_V \frac{\sigma^2}{E} dV = \frac{1}{2} \int_V s_1 (-Mz/I)^2 dV = \frac{s_1}{2I} \int_0^L M^2 dx \\ &= \frac{s_1}{2I} \int_0^L \left(2 \frac{I}{s_1}\right)^2 a^2 \sin^2(\omega t) dx = \frac{WH^3}{6s_1} La^2 \sin^2(\omega t) \end{aligned}$$

where we have used the fact that  $I = WH^3/12$ . At resonance the amplitude of kinetic and elastic energy are equal:

$$\begin{aligned} \frac{1}{10} \rho HW a^2 L^5 \omega_0^2 &= \frac{WH^3}{6s_1} a^2 L \sin^2(\omega t) \\ \Rightarrow \omega_0^2 &= \frac{5H^2}{3\rho s_1 L^4} \end{aligned}$$

The effective mass is then given by:

$$\begin{aligned} \omega_0^2 &= \frac{k}{m} \\ \Rightarrow m &= \frac{2(s_1 + s_2)\rho L^3 W}{5s_1 t} \end{aligned}$$

The last parameter of the model would be the loss  $c$  which is harder to evaluate analytically without further hypothesis. In practice, we would estimate from past experience – or measure – the quality factor  $Q$  of the structure and set the value of  $c = \frac{\sqrt{km}}{Q}$  accordingly.



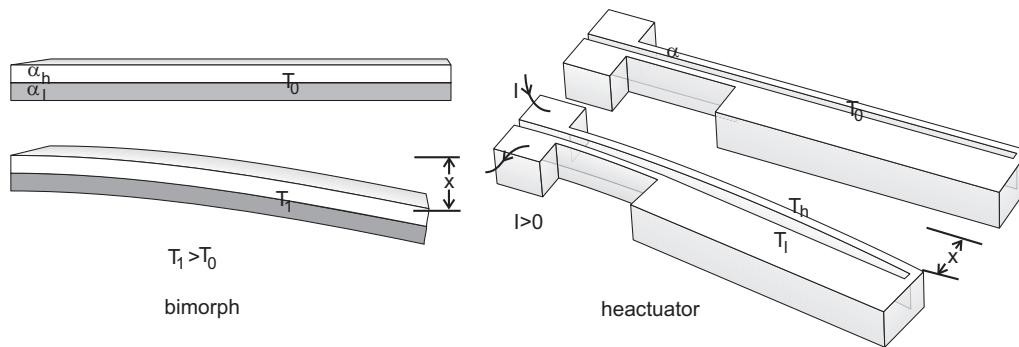


Figure 4.30: Thermal bimorph actuators.

different temperatures (e.g., polysilicon and metal) makes any bilayer curl when it is released at room temperature. This effect is often annoying, and if it can be controlled to some extent, it is the main issue behind the use of bilayer actuator. Actually, for such actuator, upon release the bilayer will curl (up if it has been well designed!) and as its temperature changes (e.g. using Joule's heating by flowing a current) its radius of curvature evolves – but it will be difficult to make it flat. Still, the initial stress induced curvature has been put to good use to fabricate curling beam that naturally protrude high above the surface of wafer in a permanent way. They have been used to lift other MEMS structure (e.g. micro-assembly in Figure 5.1) or micro-parts (e.g. conveying system).

The heatuator [20] does not have this problem as it uses a single material. This simplifies the fabrication, and the difference in strain is obtained by maintaining different temperature in the two arms. Actually as the current flow through the actuator the wider 'cold' arm will have a lower resistance and thus generate less heat than the other narrow 'hot' arm.

It should be noted that the force produced by these two actuators decreases with the stroke: at maximum stroke all the force is used to bend the actuator and no external force is produced. One heatuator can produce force in the  $10\ \mu\text{N}$  range and they can be connected together or made thicker to produce larger force.

The thermo-pneumatic actuator is another actuator where the expansion of a heated fluid can bulge a membrane and produce a large force. This principle has been used to control valve aperture in micro-fluidic components. In the extreme case, the heating could produce bubble resulting in large change of volume and allowing to produce force as in the inkjet printer head from Canon or HP.

Finally the shape memory effect is also controlled by temperature change and traditionally belongs to the class of thermal actuator. The shape memory effect appears in single crystal metal like copper and in many alloys among which the more popular are NiTi (nitinol) or  $\text{Ni}_x\text{Ti}_y\text{Cu}_z$ . In such shape memory alloys (SMA) after a high temperature treatment step, two solid phases will appear one at low temperature (martensite phase) and the other at high temperature (austenite phase). The alloy is rather soft and can be easily deformed at low temperature

in the martensite phase. However, upon heating the alloy above its phase transition temperature it will turn to austenite phase and returns to its original shape. This process creates large recovery forces that can be used in an actuator. The temperature difference between the two phases can be as low as 10°C and can be controlled by changing the composition of the alloy. In principle the alloy can be 'trained' and will then shift from a high temperature shape to a low temperature shape and vice-versa when the temperature is changed. In practice, training is difficult and micro-actuators based on SMA are one way actuator, the restoring force being often brought by an elastic member, limiting the total deformation. The most common application of such material has been for various micro-grippers, but its use remains limited because of the difficulty in controlling the deposition of SMA thin-films.

## Problems

1. Establish the expression of the spring constant of the folded beam suspension. You may want to consider the symmetry existing in the structure and decompose it as a set of cantilever beams connected in series and in parallel.
2. We consider a micro-cantilever of length  $L$ , width  $w$  and thickness  $h$  bending under its own weight.
  - What is the expression of the weight per unit of length of the cantilever assuming the material has a density of  $\rho$ ?
  - What is the general expression of the deflection at the tip of the cantilever?
  - What is the length of a  $2\ \mu\text{m}$  thick silicon cantilever whose tip deflects by  $2\ \mu\text{m}$ ? (Note: the density and Young's modulus of silicon are  $\rho = 2.33 \cdot 10^3\ \text{kg/m}^3$  and  $E = 106\ \text{GPa}$ , the acceleration of gravity is  $g=9.81\ \text{m/s}^2$ )
  - What are the practical implication of this deflection for the cantilever?
3. A force sensor is based on a piezoresistor placed on a cantilever. For which layout in Figure 4.31 will the force sensitivity be the highest when the force is applied at the tip of the cantilever? (Note: we have  $\pi_l = -31 \cdot 10^{-11}\text{Pa}^{-1}$  and  $\pi_t = -17 \cdot 10^{-11}\text{Pa}^{-1}$ .)

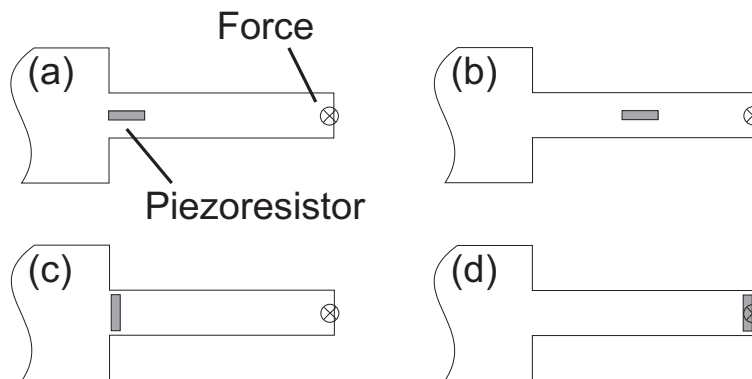


Figure 4.31: Design of force sensor.

4. A capacitive accelerometer uses the structure of Figure 4.32, where  $2L = 300\ \mu\text{m}$  and  $w = 10\ \mu\text{m}$ . The structural layer is in silicon ( $E = 130\ \text{GPa}$ ,  $\rho = 2300\ \text{kg/m}^3$ ) and  $t = 2\ \mu\text{m}$ .
  - Discuss the characteristics (bandwidth, sensitivity) and the trade-off for such accelerometer.

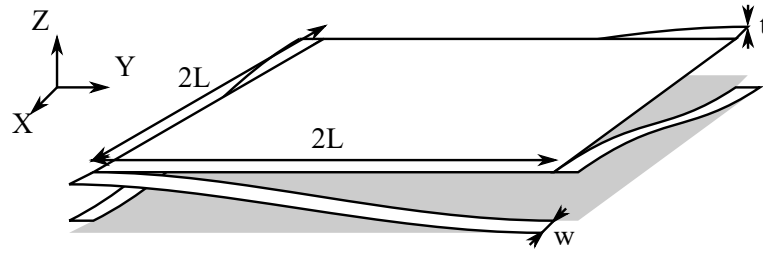


Figure 4.32: Sketch of a Z-axis accelerometer.

- Show that each spring of the suspension may be decomposed as 2 cantilevers and compute the resulting spring constant of the suspension.
  - Compute the resonant frequency of the accelerometer and deduce the bandwidth of the sensor if we want a maximum error of 1%, assuming the damping is  $\zeta = 1/\sqrt{2}$ .
  - The fabrication tolerance for the thickness of the deposited thin-film is  $\pm 0.1 \mu\text{m}$  and the in plane tolerance during photolithographic step is  $\pm 0.5 \mu\text{m}$ . What is the relative error that we can expect on the accelerometer sensitivity (we assume that  $E$  and  $\rho$  have with a much better precision) ? Which tolerance has the largest effect and what could you propose to improve this figure ?
5. A capillary of diameter  $100\mu\text{m}$  is placed vertically above a reservoir. How high will water rise in the channel if it is made in silicon (contact angle  $63^\circ$ ) or if it has been coated with silicon nitride (contact angle  $24^\circ$ )? What happens in the capillary if water is replaced by mercury (considering contact angle remain the same)?
6. Establish the equation (eq. (4.7)) for the pressure drop in the capillary shown in Figure 4.33, starting from the expression of the Young-Laplace's equation (eq. 4.6).

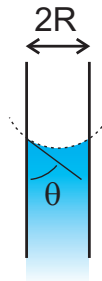


Figure 4.33: Capillary cross-section.

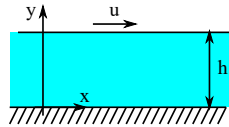


Figure 4.34: 2D channel with moving top wall.

7. We consider the channel in Figure 4.34 where the fluid is set in motion by moving the top wall at a speed  $u$  (there is no pressure difference nor acceleration along  $x$ ).
  - Where is the fluid velocity maximal ?
  - Supposing that the Stokes conditions are verified, express the velocity in the channel as a function of  $y$  and sketch the velocity profile in the channel ?
  - What is the average speed in the channel if  $u = 1 \text{ mm/s}$  ?
8. Plot the value of the  $Re_f$  factor in the Darcy-Weissbach formula (eq. (4.11) p. 187) for rectangular channel of arbitrary aspect ratio  $b/a$ .

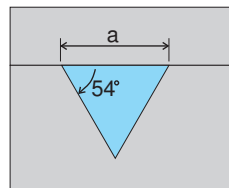


Figure 4.35: Channel cross-section.

9. A channel etched in silicon by KOH is then sealed with wafer bonding resulting in the cross-section shown in the Figure 4.35. What will be the approximate expression of the flow as a function of the pressure drop for such a channel?
10. Show that for Poiseuille's flow in a cylindrical channel of radius  $R$  the average velocity is half the maximum velocity at the center of the channel (you will need to express Stokes equation in cylindrical coordinates).

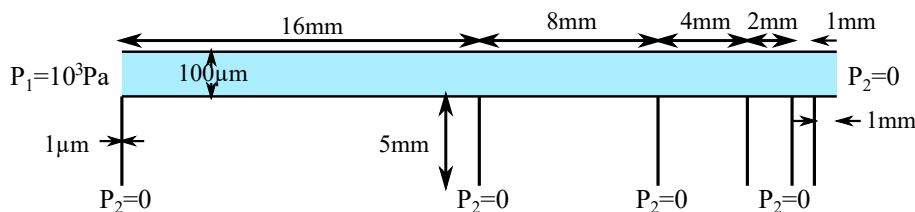


Figure 4.36: Channel network.

11. We consider the channel network shown in Figure 4.36 where a main channel of width  $100\ \mu\text{m}$  branches toward many smaller channels of diameter  $1\ \mu\text{m}$ . Water is the fluid used in the channels and we may neglect the effect of the junction geometry).
- What is the hydraulic resistance  $R$  of the main channel for a length of  $1\ \text{mm}$ ?
  - What is the hydraulic resistance  $r$  of a small channel of length  $5\ \text{mm}$ ? Compare it with  $R$ .
  - Draw the equivalent electric network for these channel network.
  - What is the flow rate for the leftmost small channel at the entrance of the main channel?
  - Neglecting some flow rate, calculate the flow rate at the output of each of the small channels.
  - If an on-off valve is placed on each small channel, what could be the role of this fluidic circuit?
12. Establish the expression of the pull-in voltage in the electrostatic gap-closing actuator shown in Figure 4.26. You can start first by writing the equilibrium equation between the restoring spring force and the electrostatic force.
13. We consider the fishbone resonant structure shown in Figure 4.37 fabricated with a conductive material with a density  $\rho = 2.7\ \text{g/cm}^3$ , an elasticity modulus  $E = 70\ \text{GPa}$  and a thickness  $t = 1.2\ \mu\text{m}$ .
- Find the expression of the effective mass of a cantilever of length  $L$
  - Find the vertical resonant frequency of this cantilever and show that it is independent of its width  $w$ . What is the length  $L$  of a cantilever with a resonant frequency of  $10\ \text{kHz}$ ?

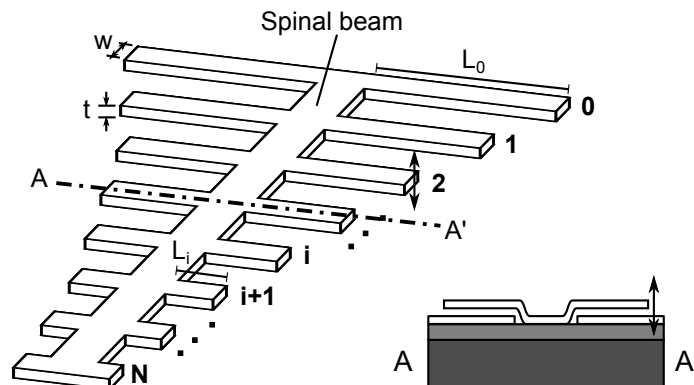


Figure 4.37: Resonant fishbone structure

- Find the horizontal resonant frequency of this cantilever (that is along  $w$ ). What should be the width of the cantilever if we want to obtain a horizontal resonant frequency at least 10 times larger than the vertical resonant frequency.
- What is ratio of the length between two consecutive cantilever  $L_{i+1}/L_i$  if we want that from one cantilever to its neighbour the vertical resonant frequency is multiplied by a factor of two ?
- The release process limits the cantilever length to  $500 \mu\text{m}$  and the linear behaviour of the cantilever imposes to keep a ration  $L/w > 5$ . What is the maximum number of cantilever  $N$  that can be fabricated with this process if there is a factor of two between the resonant frequencies of neighbouring cantilevers?
- Sketch the reduced order lumped model of this system. What element is missing to make a full simulation?

## Solutions

### Problem 8

We have seen that using the Darcy-Weissbach approximation we may write:

$$\dot{V} = \frac{2AD_h^2}{Re f \eta L} \Delta p$$

and that the exact solution of the Stokes equation for rectangular cross-section gives:

$$\dot{V} = -\frac{dp}{dx} ab \frac{64}{\eta \pi^6} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}$$

thus we see that  $\frac{2D_h^2}{Re f} = b^2 \frac{64}{\pi^6} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}$ .

For a rectangular cross-section we have  $D_h = \frac{2ab}{a+b}$ , that we may introduce in the previous equation, giving  $\frac{8a^2}{Re f (a+b)^2} = \frac{64}{\pi^6} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}$  and finally we write:

$$Re f = \frac{8}{\left(1 + \frac{b}{a}\right)^2 \frac{64}{\pi^6} \sum_{n=2,4,\dots}^{\infty} \sum_{m=2,4,\dots}^{\infty} \frac{1}{n^2 m^2 ((b/a)^2 n^2 + m^2)}}$$

and plot the corresponding result as a function of  $\epsilon = b/a$  using a numerical software to compute the infinite sum:

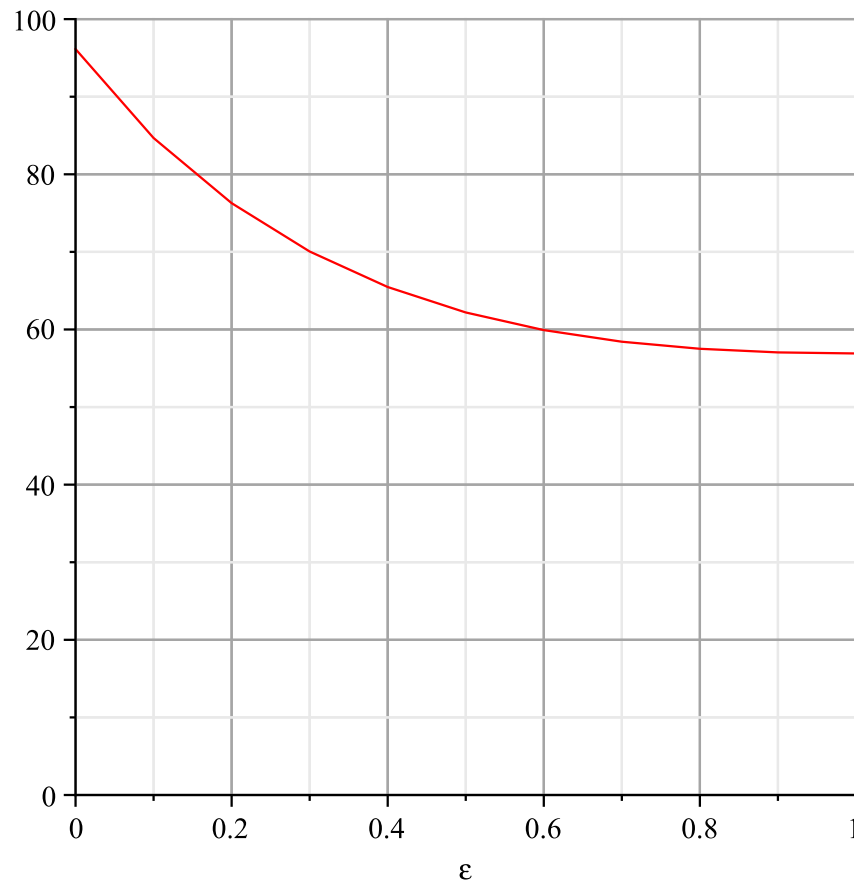


Figure 4.38: Variation of the  $Ref$  factor in the Darcy equation for a rectangular channel with varying aspect ratio  $\epsilon = b/a$ .

There is in the literature another way of presenting the Darcy equation as a function of hydraulic diameter, using the area of the circular channel instead of  $A$ , the cross-section area of the real channel:

$$\dot{V} = \frac{\pi D_h^4}{2Ref\eta L} \Delta p$$

It clearly gives the same results for circular channel for  $Ref$  but it does not give the right geometry dependency ( $b^4$  instead of  $ab^3$ ) for the slab channel and the expression using  $A$  should be preferred.

### Problem 10

For incompressible fluid with laminar flow the Navier-Stokes equations can be simplified and becomes the Stokes equation (4.8):

$$\overrightarrow{\text{grad}} p = \eta \Delta \vec{u} + \vec{F}$$



where  $\vec{F}$  is a body force that applies everywhere on the liquid that we will ignore here. Flow needs also to verify the mass conservation equation written as

$$\operatorname{div} \vec{u} = 0$$

and expressing that the net flow at any point is 0 (i.e. as much fluid entering than leaving).

We consider flow inside a cylindrical channel (with circular cross-section), neglecting body force, and consider we have a Stokes flow. Because of the geometry we will use cylindrical coordinates and try to find  $\vec{u}(u_\rho, u_\varphi, u_z)$ , after considering some simplifying arguments.

First, as the flow is laminar the streamlines are only along the length of the tube, that is  $\vec{u} = u_z \hat{\mathbf{z}}$ , thus  $u_\varphi = u_\rho = 0$  and we only need to solve a simpler scalar function ( $u_z$ ) to obtain  $\vec{u}$ .

Then we use the conservation equation and write in cylindrical coordinates:

$$\operatorname{div} \vec{u} = \frac{1}{\rho} \frac{\partial \rho u_\rho}{\partial \rho} + \frac{1}{\rho} \frac{\partial u_\varphi}{\partial \varphi} + \frac{\partial u_z}{\partial z} = 0$$

replacing the value of  $u_\varphi = u_\rho = 0$  we find easily that  $\frac{\partial u_z}{\partial z} = 0$ , thus the velocity is independent of  $z$  and conserved along the tube (another way to look at it is to say that there is the same flow rate – as the section of the tube is constant – at the input and output of the tube).

Then we will consider the right side of the Stokes equation. The Laplacian of the velocity vector is expressed in cylindrical coordinate and simplified knowing that  $u_\varphi = u_\rho = 0$  as:

$$\begin{aligned} \Delta \vec{u} &= \left( \Delta u_\rho - \frac{u_\rho}{\rho^2} - \frac{2}{\rho^2} \frac{\partial u_\varphi}{\partial \varphi} \right) \hat{\rho} \\ &\quad + \left( \Delta u_\varphi - \frac{u_\varphi}{\rho^2} + \frac{2}{\rho^2} \frac{\partial u_\rho}{\partial \varphi} \right) \hat{\varphi} + \Delta u_z \hat{\mathbf{z}} \\ &= \Delta u_z \hat{\mathbf{z}} \end{aligned}$$

We see that the vectorial Laplacian became a simpler scalar Laplacian, whose expression in cylindrical coordinate is:

$$\Delta \vec{u} = \Delta u_z \hat{\mathbf{z}} = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u_z}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 u_z}{\partial \varphi^2} + \frac{\partial^2 u_z}{\partial z^2} \hat{\mathbf{z}}$$

however, for symmetry reason, the velocity  $u_z$  is clearly independent of the angle  $\varphi$  and, as conservation law has just shown above, independent of  $z$ , thus it depends only on  $\rho$  and we may simplify the equation to:

$$\Delta \vec{u} = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u_z}{\partial \rho} \right) \hat{\mathbf{z}}$$

If we look now at the left member of the Stokes equation, the gradient is expressed in cylindrical coordinate as:

$$\vec{\text{grad}}p = \frac{\partial p}{\partial \rho} \hat{\rho} + \frac{1}{\rho} \frac{\partial p}{\partial \varphi} \hat{\varphi} + \frac{\partial p}{\partial z} \hat{z}$$

The pressure is uniform across the input and output sections and the streamline are straight as we are in laminar flow, thus we may consider that the pressure depends only on the  $z$  coordinate  $p = p(z)$ , that is:

$$\vec{\text{grad}}p = \frac{\partial p}{\partial z} \hat{z}$$

. thus, the pressure gradient is only oriented along  $\hat{z}$ .

Merging the expression we found for the left and right side of the equation, we may write the Stokes equation along  $\hat{z}$  as:

$$\frac{\partial p}{\partial z} = \eta \left[ \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u_z}{\partial \rho} \right) \right]$$

that we write as:

$$\frac{\partial p}{\partial z} \frac{\rho}{\eta} = \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u_z}{\partial \rho} \right)$$

We integrate once w.r.t.  $\rho$  giving:

$$\frac{\partial p}{\partial z} \frac{\rho^2}{2\eta} = \rho \frac{\partial u_z}{\partial \rho} + A$$

where  $A$  is an integration constant. Knowing that the velocity is zero at the tube wall (no-slip boundary) it seems reasonable for symmetry reason that the velocity is maximal at the centre  $r = 0$  of the tube, thus  $\frac{\partial u_z}{\partial \rho}(0) = 0$  and we have in this point:

$$0 = 0 + A$$

thus  $A$  is null. We divide both side by  $\rho$  ( $\rho > 0$ ), giving:

$$\frac{\partial p}{\partial z} \frac{\rho}{2\eta} = \frac{\partial u_z}{\partial \rho}$$

and integrate a second time w.r.t  $\rho$ :

$$\frac{\partial p}{\partial z} \frac{\rho^2}{4\eta} + B = u_z$$

Using again the no-slip boundary condition, we have  $u_z(R) = 0$  and:

$$B = -\frac{\partial p}{\partial z} \frac{R^2}{4\eta}$$

Finally we may write  $u_z$  as :

$$u_z = \frac{\partial p}{\partial z} \frac{R^2}{4\eta} \left( \frac{\rho^2}{R^2} - 1 \right)$$

The velocity profile across the tube is parabolic, which is a clear signature of Poiseuille's flow.

In this channel the maximum velocity is at the center in  $\rho = 0$  and is expressed as:

$$u_{\max} = u_{z\max} = -\frac{\partial p}{\partial z} \frac{R^2}{4\eta}$$

Moreover the average velocity is given by:

$$\begin{aligned} \bar{u}_z &= \frac{1}{\pi R^2} \int_0^R \int_0^{2\pi} \frac{\partial p}{\partial z} \frac{R^2}{4\eta} \left( \frac{\rho^2}{R^2} - 1 \right) \rho d\varphi d\rho \\ &= \frac{\partial p}{\partial z} \frac{1}{4\pi\eta} \int_0^R \int_0^{2\pi} \left( \frac{\rho^2}{R^2} - 1 \right) \rho d\varphi d\rho \\ &= \frac{\partial p}{\partial z} \frac{1}{2\eta} \int_0^R \left( \frac{\rho^2}{R^2} - 1 \right) \rho d\rho \\ &= \frac{\partial p}{\partial z} \frac{1}{2\eta} \left( \frac{R^2}{4} - \frac{R^2}{2} \right) \\ &= -\frac{\partial p}{\partial z} \frac{R^2}{8\eta} \end{aligned}$$

and we finally have  $\bar{u} = \bar{u}_z = \frac{1}{2}u_{\max}$ , as expected. We may note that this is different from the 2D channel case ( $\bar{u} = \frac{2}{3}u_{\max}$ ), and it can be expected that rectangular channel will have a factor depending on their aspect ratio : if they are more circular it will be near  $\frac{1}{2}$  and if they are much more wide than deep, it would be near  $\frac{2}{3}$ .



## Chapter 5

# MEMS packaging, assembly and test

MEMS packaging, assembly and test – collectively called back-end process – problems are the aspects of the MEMS technology that even now remain the less mature. Actually, although the bookshelves appear to be replete with books discussing all the aspects of MEMS technology, we had to wait until 2004 to finally have a reference book really discussing these three issues with real MEMS examples [29]. However it is hard to stress enough how MEMS packaging and test are important for obtaining a successful product at a low cost. Figure 5.1 shows two real life examples of MEMS based micro-sensor where packaging, assembly and test are a dominant part of the total cost.

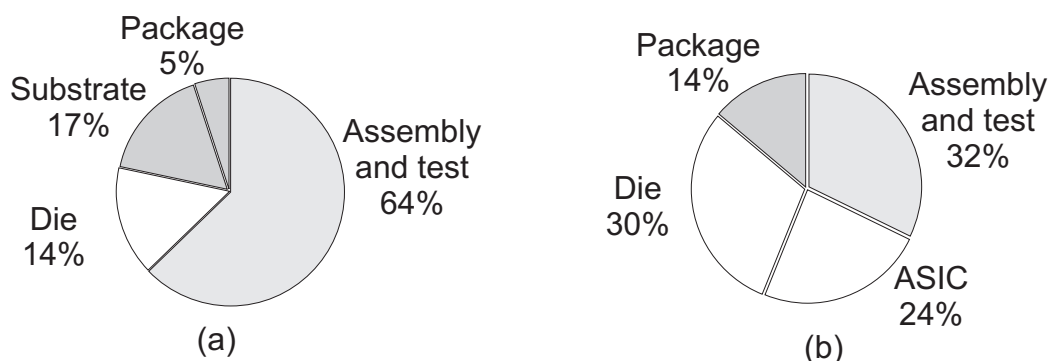


Figure 5.1: Cost break-up in (a) a pressure sensor in plastic package (b) an accelerometer in surface-mount package

As a matter of fact, the choice taken for packaging and test may dictate how to design the MEMS chip itself! This is in sharp contrast with micro-electronics packaging where packaging is a somewhat independent activities than chip processing or design. On the contrary, in the MEMS case (no pun intended!), the influence of the package on the microsystem behavior may be very significant.

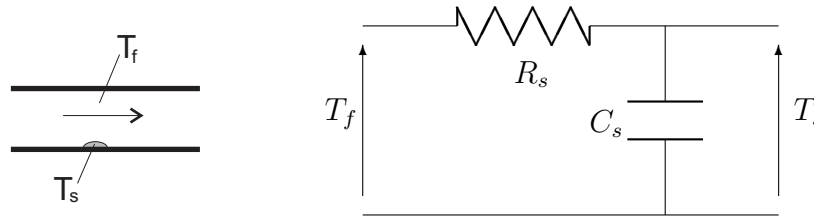
---

**Example 5.1** Packaging affecting MEMS response
 

---

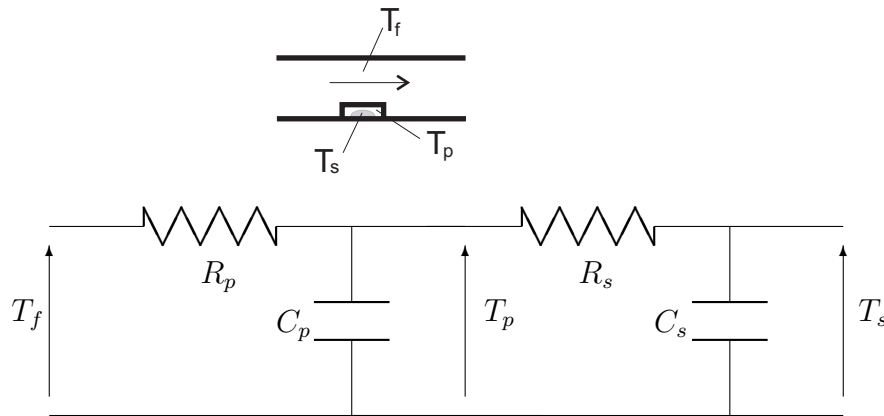
WE CONSIDER A THERMAL SENSOR, used in a flow sensor and look at two different scenarios for the packaging : either the sensing element is directly in contact with the fluid or, because the fluid is corrosive, it is protected by a sheath.

In the first case the sensor is a first order system, where  $R$  is the thermal resistance (the way it oppose change in temperature) of the sensing element and  $C$  its thermal capacitance (how it stores heat).



Thus its transfer function is given by:  $\frac{T_s}{T_f} = \frac{1}{1+\tau_s s}$  with  $\tau_s = R_s C_s$

However, when a sheath is placed around the sensing element it brings an additional thermal resistance ( $R_p$ ) and thermal capacitance ( $C_p$ ), and the sensor becomes a second order system.



with the transfer function :

$$\frac{T_s}{T_f} = \frac{1}{1 + (\tau_p + \tau_s + \tau_s R_p / R_s) s + \tau_s \tau_p s^2} \text{ with } \tau_p = R_p C_p \text{ and } \tau_s = R_s C_s$$

The dramatic differences existing between responses of first and second order systems clearly underscore the necessity to include the packaging at an early stage of the design.

---

The Example 5.1 shows clearly that early consideration of the packaging solution could lead to a successful product... while ignoring it could lead to a dramatic failure.

The different operations that could appear in a somewhat complete MEMS back-end process are presented in Figure 5.2, and we find the assembly, packaging and testing steps. However as we stressed earlier, nothing is less typical than a ‘standard’ MEMS back-end process - and in practice, some steps may not be present or their order be different.

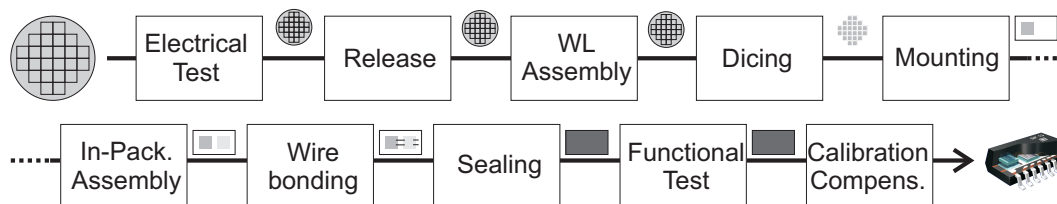


Figure 5.2: Standard packaging steps for MEMS

The general tendency, to decrease the cost, is to perform as many steps as possible at the wafer level and not for individual die. Accordingly, recent years have seen the emergence of techniques like wafer Level Packaging (WLP), where the back-end process can hardly be distinguished with the front-end process. However, in most cases, many operation still need to be performed component by component, explaining largely the cost of this step in the MEMS production.

## 5.1 Assembly

Assembly of MEMS part is normally not a good idea as it traditionally requires part-by-part processing and loose the cost advantage attached to batch processing.

However there are examples that show it can still be used and we have discussed previously (4.1) that hybrid integration is the normal answer to electronics integration with MEMS. The IC chip is assembled with the MEMS chip either quasi-monolithically using flip-chip solder bump, or more simply using in-package assembly (Figure 3.23) : the MEMS part is placed inside the package along the electronics chip and connected using wire-bonding. Such systems are referred to SIP (System In the Package) and contrast strongly with the SOC (System On Chip) approach, where MEMS and electronics are built together on the same die like in the ADXL accelerometer from Analog-Devices.

Assembly is sometimes a viable solution, and it is actually even possible to have real MEMS part assembly. After all, watch-makers assemble small parts and still manage to reach very low cost using automation. Accordingly, different automated MEMS assembly systems have been demonstrated, and some showed success for tasks with reduced manipulation need. However complete free 3D assembly of MEMS part (6DOF) with the necessary accuracy and precision ( $< 1\mu\text{m}$ ) is still

too slow to be of practical use.

However MEMS offers a much clever path to assembly than serial processing: the batch assembly, which maintains the advantage of batch processing. This is accomplished along two paths : self-assembly or integration of assembly micro-actuator.

In the first case, natural forces like capillary forces, are used to pull and position the MEMS element in place.

In the second approach, at the same time the MEMS is developed, micro-actuator are integrated to realize the assembly task. In this way, we can recoup some of the overhead brought by assembly, because now, the mechanical assembly can be performed by batch. Actually for allowing complete assembly, the MEMS should not only include an assembly actuator but also some locking structure that will keep the assembly in position even after the special actuators are no more powered. As they are used only for a short time, these actuators can have a short lifespan (e.g., SDA actuator) even working one time only (e.g., stress based actuator as in Example 5.2), require large power (e.g., heatuators) or high voltage (e.g. large force comb-drive actuator) without posing too much problem.

## 5.2 Packaging

MEMS packaging, unlike the well-established and standardized IC packaging technology, is still largely an ad-hoc development. The main packaging efforts have been conducted within MEMS manufacturer companies, and they have jealously kept their secret considered, with reason, as the most difficult step to bring MEMS to market.

Still, the purpose of packaging in MEMS is in many ways similar to IC, and we can list a series of functionalities that should be brought by the package:

- Support: provide a standard mechanical support for handling during assembly of the MEMS part into a system.
- Protection: protect the chip from the environment (dust, stress, shock, moisture...). For MEMS the most important parameter to be controlled is often stress.
- Interfacing: bring signal in and out of the chip. For MEMS, signals will not only be electrical but may be fluids, radiation, fields...
- Heat removal: ensure the heat generated inside the chip is properly evacuated to the environment (it is actually only a special kind of interfacing, but not for a signal). Actually, this point is much less severe with MEMS than with IC and is usually relatively easy to fulfill.

The need for protection from the environment is actually for MEMS not only for reliability (e.g. preventing corrosion of metal contact) but it often serves to



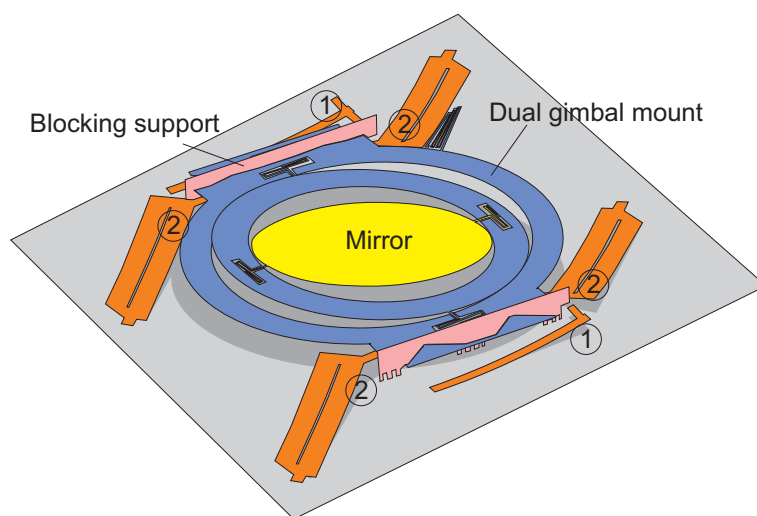
---

**Example 5.2** Stress-driven automated assembly of tunable micro-mirror.
 

---

ALCATEL-LUCENT BELL LABS did develop at the turn of the millennium (and in 18 months!) the Lambda router (marketed until mid 2002 as the core component in their Wavestar all-optical routing system) using an intricate structure of integrated stress-driven vertical actuator to assemble array of two-axis gimbal mirrors.

The complex mirror structure needed a fast electrostatic command that could be easily obtained with surface micromachining. However they also needed a large tilting angle that this process could not directly provide as deposited sacrificial layer are too thin to obtain enough space below the 500  $\mu\text{m}$  wide mirrors. They had to develop a clever assembly process to obtain from the flat surface-micromachined layers the 3D structures they needed.



The mirror is actually pushed above the surface of the wafer using beams curling up permanently under stress gradient (see Section 4.5.4) and held in operating position by additional blocking structure. The Figure shows the principle of the device which get automatically assembled during the release etch: at first (①) the narrow bilayer beams get freed positioning the locking structure above the gimbal mount 'ears', then (②) the wider bilayer beams are freed, curling up and pushing the mirror against the blocking structure that self-align using the V-shaped frames.

---

solve simultaneously a functional problem. For example, hermetic encapsulation protects die from contamination but it is also used in pressure sensor to obtain a reference pressure in a cavity, or to control damping in resonant system. But the required protection level in MEMS is often higher than for IC. Actually the presence of water vapor - that could condensate during use on any mobile part - may make hermetic package a must have and not only for pressure sensor. For example, a water droplet appearing inside the TI's DMD mirror array would induce

defects hard to accept in a product of that price.

To understand some of the challenge that lies in the design of MEMS packaging we may have a look at the Table 5.1 adapted from [28], where, next to some typical MEMS micro-sensors, we have also figured the requisite for micro-electronics packaging.

Sensors	Elec. port	Fluid port	Transp. wind.	Hermetic encaps.	Stress isolation	Heat sink	Thermal isolation	Calib. Comp.
Pressure	yes	yes	no	maybe	yes	no	no	yes
Flow	yes	yes	no	no	no	no	yes	yes
Accel.	yes	no	no	yes	maybe	no	no	yes
Yaw-rate	yes	no	no	yes	maybe	no	no	yes
Sound	yes	yes	no	no	no	no	no	yes
Light	yes	no	yes	no	no	no	maybe	no
Temp.	yes	maybe	no	maybe	no	maybe	maybe	yes
IC	yes	no	no	moisture	no	maybe	no	no

Table 5.1: Packaging and testing requirement for some micro-sensors (adapted from [28]).

It is obvious that the challenges brought by micro-sensors packaging are completely different from those encountered by electronics packaging. The necessity for the measurand to reach the sensing element brings in transparent windows, fluid port, gas hermetic sealing, stress isolation... unheard of in the IC industry, and unfortunately, in the IC packaging industry.

In fact, besides protection, the major hurdle often rests in interfacing to the external environment. Actually this problem is diametrically opposed to the preceding point: interfacing requires us to open a way in (and out) through the protection to come close to the MEMS die. For inertial sensors, such as accelerometers and gyroscopes, the packaging problem is not too severe because they can be fully sealed and still sense the measurand they are to probe provided they are rigidly attached to the package. In that case, the use of stress relieving submount and a bonded cap is all what's needed to be able to use modified IC packaging procedure. But this is for the simplest cases, for chemical and biological sensors, which must be exposed to fluids, the task is much more complex and the package can represent as much as 90% of the final cost. Actually, the diversity of issue encountered for interfacing has for the moment received no standard solution and the packages are then designed on a case-by-case basis. In many cases the package will condition the response of the MEMS, particularly in the case of micro-sensor, and the package must be considered during design at the earliest possible stage. See for example the gas sensors developed by Microsens in Figure 5.3. The package use a charcoal filter placed inside the cap for a very important function : decreasing cross-sensitivity by allowing only small molecule gas to go through and to reach the sensing element. The time for the gas to diffuse through the filter determines mostly the response of the sensor, that behaves as a first order system with a time

constant in the order of 10 s, whereas the response of the sensing element (the ‘MEMS’ die) is shorter than 1 s. The package has definitely a dramatic effect on the system response!

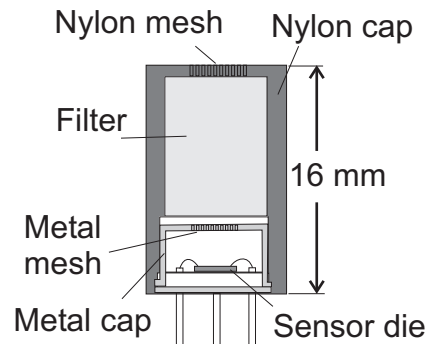


Figure 5.3: A gas sensor and its package developed by ©Microsens

### 5.2.1 Encapsulation

Encapsulation is used to protect the MEMS from mechanical contact, dust, water vapour or other gases that could affect the reliability of the MEMS. Encapsulation is performed by using plastic, ceramic or metal, with cost increasing in the same order. For example, a pressure sensor from Novasensor packaged in plastic may be sold for less than US\$5, while a steel housed sensor with metallic membrane for harsh environment may exceed the US\$100 mark! NovaSensor is proposing the three types of packaging: the plastic package for a low-cost low performance sensor, a ceramic package for compensated structure for medical application, a metal case (modified from a standard TO-8 cap from the early IC industry) and a full metal package (the interface to the environment is a metallic membrane) for corrosive or harmful media. But we should note that the MEMS sensing element used in these three sensors is exactly the same - the value of the sensor is mostly in the package!

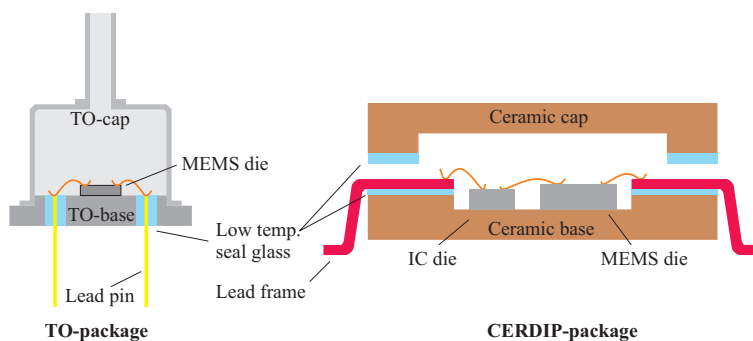


Figure 5.4: Box and lid package in metal (TO-header) and ceramic (CERDIP).

Material	CTE ppm/K	Modulus (T/S) GPa	Yield strength (T/S) MPa
Si	4.1	150	N/A
Al (pure/alloy 1100)	23	69	35/60
Al (Si alloy)	6.5-13.5	100-130	
Cu	17	117	70
Ni	13	207	148/359
In	29	11	1.9/6.1
Alumina (Al <sub>2</sub> O <sub>3</sub> )	6.7	380	
Kovar	5.9	131	340
Borosilicate glass	4.5	65/26	29
Epoxy (pure)	60	2.4/0.9	54
Epoxy (carbon fiber)	-1.1	186	

Table 5.2: Characteristics of some packaging materials (T:tensile, S:shear).

As MEMS generally have mobile parts, the main difficulty of MEMS encapsulation is to avoid blocking this motion. This originally forbid the simple use of injection molding after mounting on a lead frame as in standard IC packaging. The earlier idea were to use existing ‘box and lid’ type of casing, like the older TO series of package used for transistors or the ceramic Cerdip package, as we see in Figure 5.4. This package have been adapted to the specific MEMS application, and for example we have figured an original Transistor Outline (TO) case modified by welding a fluid entry port to the cap for using it in a pressure sensor. Later on, to reduce cost, injection molded plastic ‘box and lid’ cases were used, most notably in Motorola’s consumer grade range of pressure sensors. The Cerdip case solves the second main issue generally encountered in MEMS packaging : thermal stress. Many MEMS devices are actually stress sensitive, and besides the obvious piezo-resistive type of sensors, even DLP chips or fluidic micro-valves would have issue if too much stress is induced during the encapsulation step or during the device use. As such, the choice of materials is often tied to the coefficient of thermal expansion (CTE) with respect to silicon. The Alumina used in Cerdip package has a CTE close to the CTE of silicon and thus won’t introduce too much stress changes while it is operated in the environment.

Another example of matched CTE in MEMS packaging is given by the DLP. Actually the glass window above the chip is made in borosilicate glass that is bonded to a kovar lid. A rapid look in Table 5.2 shows that this choice of material

is well thought of: they have both a matching CTE which is close to the CTE of silicon!

However, matching the CTE is not the only way a material may be suitable for MEMS packaging - particularly for a material that is in contact with the MEMS die. Actually some polymer, although they exhibit a very large CTE and will experience large thermal expansion, can still be used for packaging. To understand that, we note that the stress induced by thermal expansion is proportional to the temperature  $T$ , the CTE  $\alpha$  but also to the Young's modulus  $E$ :

$$\sigma_T \propto E\alpha T$$

As polymer have usually small Young's modulus, they won't exert much force on the silicon die as they will deform readily and absorb much of the induced strain. In fact soft polymer are often used as a stress relieving buffer to attach the die to the casing for example.

More surprisingly, some material, that possess a high CTE and a relatively high Young's modulus can still be used in packaging. The best example is given by indium. This metal can be used for solder or as a paste for die attachment because, although it has a relatively large CTE and Young's modulus, its yield strength is very low. Thus as the thermal strain changes, the alloy will quickly deform plastically and again won't induce any excessive stress on the MEMS die.

If MEMS encapsulation is still too often an adhoc development, some strategies are maturing to keep as much as possible of the cost advantage brought up by batch fabrication. As such, more and more MEMS devices use first level encapsulation, where a glass or silicon wafer is bonded to the chip, helping to maintain the MEMS integrity during dicing and further mounting in the package. In the packaging with wafer bonding technique (cf. Sec. 3.4.3) the cap wafer is first patterned with a simple cavity, or even a hole if an access to the environment is needed. Then, alignment of the MEMS wafer and the cap wafer brings the cavity in front of the MEMS part before the bonding is finally performed (Figure 5.5). This

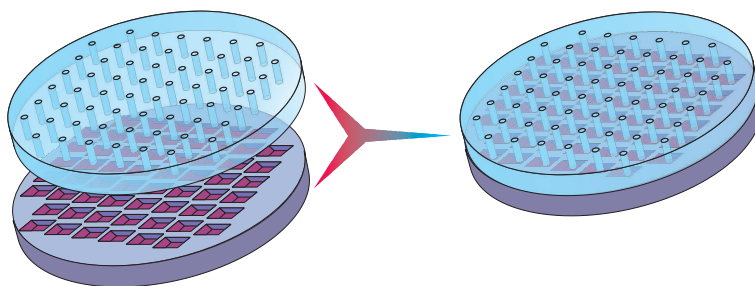


Figure 5.5: Glass to silicon anodic bonding to provide first-level encapsulation of MEMS sensor.

pre-encapsulation technique has the advantage to be wafer level and to provide an inexpensive way to allow the use of the epoxy overmolding to package the

component. Actually, after dicing the wafer in chips, the capped MEMS is rather sturdy and can be processed using standard IC packaging procedure.

The first step, shown in Figure 5.6, consists in placing each die on a lead frame, a long strip of identical metal structures punched from a thin foil. The lead frame is used as a support for the die and to obtain electrical connection that can be soldered on a printed circuit board (PCB). The dies are first glued on each of the die bonding site using polymer or indium. Then electrical wiring is done to connect the pad on the MEMS chip to the lead frame contacts that reroutes them to the chip contacts. This part of the process is serial in nature, but can benefit heavily from automation (pick-and-place and wire-bonding machines) as it is a simple task, making it surprisingly cost effective.

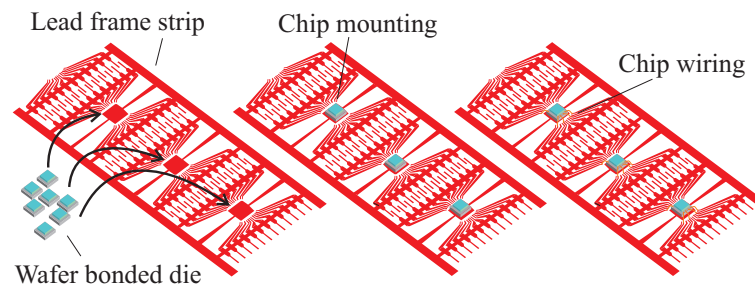


Figure 5.6: Die mounting and wiring on a lead frame.

The encapsulation steps is, in the other hand, done by batch, using directly the lead frame strip as shown in Figure 5.7. The lead frames with their wired dies are placed in a mold having multiple cavities before injection molding can be used. A thermosetting monomer ‘puck’ is heated above its glass transition temperature and pushed by a piston in all the cavities of the mould, completely covering the dies. Then heating is maintained until the thermosetting monomer is fully cross-linked and becomes a hard polymer. The polymer normally used is based on epoxy, although it has relatively poor thermal properties and needs to incorporate different additives (flame retardant for example) to meet environmental regulations.

## 5.2.2 Hermetic encapsulation

The protection function of the package is normally understood as relatively simple to ensure if one can make the package fully hermetic: in that case no contaminant (gas, moisture, dust...) can contact the die and it will be protected. However we know that even the most airtight metal box of cookies eventually get them spoiled, thus what does ‘hermetic’ really mean? What kind of leak rate may be acceptable? An easy answer will be to look at standard and for example the military standard Mil-Std-883 splits leaks between gross leak (for leak rate between  $10^{-1}$  and  $10^{-5}$  atm-cc/s) and fine leak (for leak rate between  $10^{-6}$  and  $10^{-8}$  atm-cc/s) and a gray zone in between. The standard also gives a lower limit of  $10^{-8}$  atm-cc/s for what

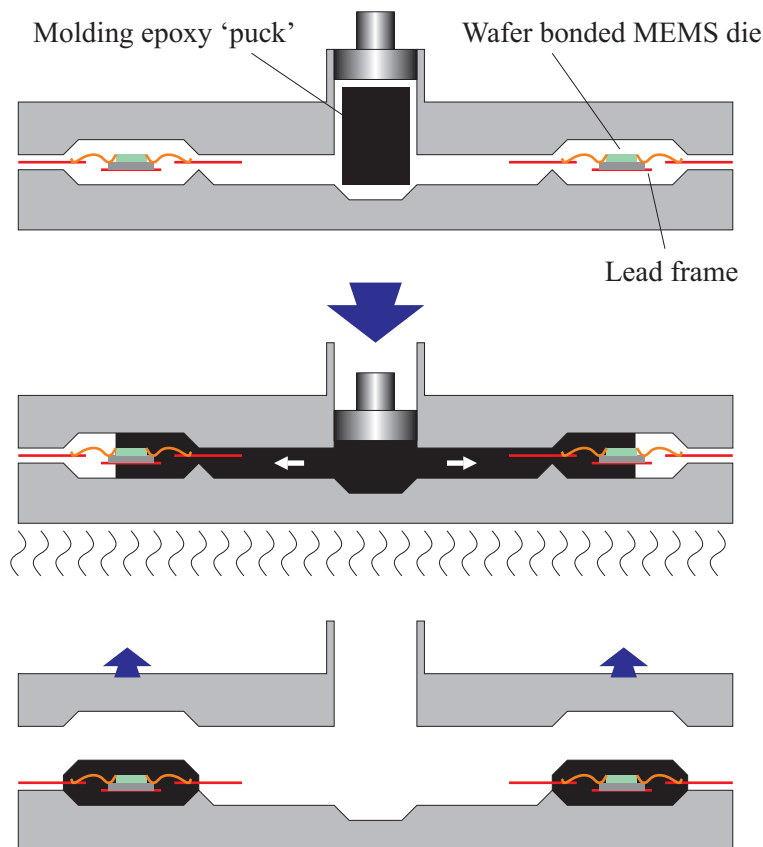


Figure 5.7: Main steps in overmolding process using thermosetting polymer (epoxy).

is required for reliable semiconductors operation – but is this OK for MEMS ? But before we try to answer this question, what actually does atm-cc/s mean?

The main characteristic of a leak is its leak rate, that describes the quantity of fluid that can pass through it in a given time. We note that the literature for leak or flow often uses a unit of pressure  $\times$  volume (at a certain temperature) instead of mol or mass for the quantity of matter – hence the at-cc/s, that is atmosphere times cubic centimeter per second. The idea behind this is that the ideal gas law directly relates  $PV$  to moles as  $PV = N_M RT$  where  $N_M$  is the number of moles. But this customary unit is ambiguous (in addition to not being a SI unit) as the quantity of matter represented by the pressure  $\times$  volume units depends on the experimental temperature  $T$  which is often not reported, making precise comparison between different systems impossible. If the temperature is known, the ideal gas law allows conversion of the customary units to mol/s as shown in Table 5.3 by using SI units of pressure and volume and dividing by  $RT$  as :

$$\frac{dN_M}{dt} = \frac{1}{RT} \frac{dPV}{dt}$$

We will be using exclusively the mol/s unit to describe leaks and encourage

Unit	0°C mol/s	20°C mol/s	25°C mol/s	300K mol/s
1 atm cc/sec	$4.46 \cdot 10^{-5}$	$4.16 \cdot 10^{-5}$	$4.09 \cdot 10^{-5}$	$4.06 \cdot 10^{-5}$
1 Pa m <sup>3</sup> /s	$4.40 \cdot 10^{-4}$	$4.10 \cdot 10^{-4}$	$4.03 \cdot 10^{-4}$	$4.01 \cdot 10^{-4}$
1 mbar l/s	$4.40 \cdot 10^{-5}$	$4.10 \cdot 10^{-5}$	$4.03 \cdot 10^{-5}$	$4.01 \cdot 10^{-5}$

Table 5.3: Conversion of flow units between customary units and mol/s for different temperatures.

engineers to do the same. Of course, the mol being representative of a quantity of matter, it can be directly linked with mass by using the molar mass  $m_M$  of the fluid ( $m = N_M m_M$ ). Accordingly, the leak rate in mol/s is also a mass flow rate, and we have  $\frac{dm}{dt} = m_M \frac{dN_M}{dt}$ . For example a leak rate of  $4 \cdot 10^{-5}$  mol/s of oxygen ( $m_M = 32$ g/mol) correspond to a mass flow rate of  $1.28 \cdot 10^{-6}$  kg/s. All the formula derived in the following using molar flow rate  $\dot{Q}_M$  could then be converted to mass flow rate  $\dot{Q}_m$  by multiplying by  $m_M$ .

Fully hermetic package, that is a package without any exchange with the environment for any period of time, hardly exists at all. In practice, given enough time, some gases will creep through some defects by diffusion like process and reach the inside of the package. Of course the existence of leaks, as can be found at seal interface, makes the thing worse, but plain materials without cracks will anyhow let fluids sip in through a process called permeation. Then the choice of the material is crucial to obtain good hermetic package, as the permeation of gas and moisture through the material itself will be the ultimate limit to the leak rate in any package.

Actually, the flow of gas  $\dot{Q}$  through a barrier made of a certain material can be linearized and using the unit of mole per unit of time ( $\dot{Q}_M = \frac{\partial N_M}{\partial t}$ ) given the form:

$$\dot{Q}_M = P_0 \frac{A \Delta P}{d}$$

where  $P_0$  is the intrinsic permeability for the material,  $A$  the exposed surface,  $\Delta P$  the pressure difference between both side, and  $d$  the barrier thickness. There is no standard unit for  $P_0$  and it mostly changes with the pressure unit used<sup>1</sup> and we use mol s<sup>-1</sup> m<sup>-1</sup> atm<sup>-1</sup>. The expression of the moisture evolution with time in a package of volume  $V_{in}$ , is obtained by first recognizing in the case of a closed volume the relationship between the flow and the pressure inside the volume using the ideal gas law ( $PV = N_M RT$ ),

$$\dot{Q}_M = \frac{dN_M}{dt} = \frac{V_{in}}{RT} \frac{dP_{in}}{dt}$$

<sup>1</sup>We express flow of matter in mol/s but the literature often reports a mass flow in kg/s instead. Divide the kg/s value by the gas molar mass (e.g. 0.018 kg/mol for water) to obtain mol/s.



then we use the definition of the permeability (and the fact that  $P_{\text{out}}$  is constant) to obtain,

$$\frac{dP_{\text{out}} - P_{\text{in}}}{dt} = -\frac{RT}{V_{\text{in}}} P_0 \frac{A}{d} (P_{\text{out}} - P_{\text{in}}).$$

This differential equation is solved by considering that the external water vapour pressure is constant ( $P_{\text{out}}(t) = P_{\text{out}} \forall t$ ), while the inside pressure is initially  $P_{\text{in}}(0) = 0$ , giving:

$$P_{\text{in}}(t) = P_{\text{out}} \left( 1 - e^{-\frac{RT}{d} \frac{A}{V_{\text{in}}} t} \right)$$

To get some general information out of this equation with a large number of parameters, we need to make a few assumptions. We will consider the time ( $t = t_{50}$ ) it takes in a cubic (spherical) box of side (diameter)  $a$  (i.e., in both cubic and spherical cases  $V_{\text{in}}/A = a/6$ ) for the water vapour pressure to reach 50% of the outside pressure. Thus using  $P_{\text{in}} = 0.5P_{\text{out}}$  we get an expression relating the permeability with other environmental and package parameters:

$$\log d - \log t_{50} + \log \left( \frac{\ln 2 a}{RT 6} \right) = \log P_0$$

Finally, considering a temperature of 25°C, and a package with a side  $a = 1$  mm, we plot Figure 5.8 showing the relationship between the time it takes to cross a certain thickness of a barrier made of materials typically used in packaging.

As a matter of fact, for electronic circuit high reliability is obtained even with non-hermetic packaging (polymer) by coating the surface of the wafer with protecting layer (e.g., low stress silicon nitride and other dielectrics) that acts as excellent barrier to moisture and gas. We can see in Figure 5.8 that 2  $\mu\text{m}$  of such materials is theoretically equivalent to almost 10 cm of epoxy! Of course the reality will be different as thin films tend to have more defects, but nevertheless the figure allows to highlight a big difference with MEMS packaging: in general mobile part will prevent using protection layers deposited on the die and the reliability will be harder to guarantee as *it will more directly depend on the hermeticity of the package itself*.

To assess and model this issue, and complete the permeability with a study of fine leaks, we will develop a model of vapour or gas leak where we describe the flow of gas with classical theory using SI units of mol/s.

The leaking behavior of packages can be described relatively simply by considering leaks as very narrow channels. Actually when there is a difference of pressure between two sides of a barrier it also means that there is a difference of gas molecules density (and maybe velocity if they are not at the same temperature) on both sides. The relationship between gas molecule molar density and pressure is simply given by using again the ideal gas law :

$$PV = N_M RT \Rightarrow P = \frac{N_M}{V} RT \Rightarrow P = n_M RT \quad (5.1)$$

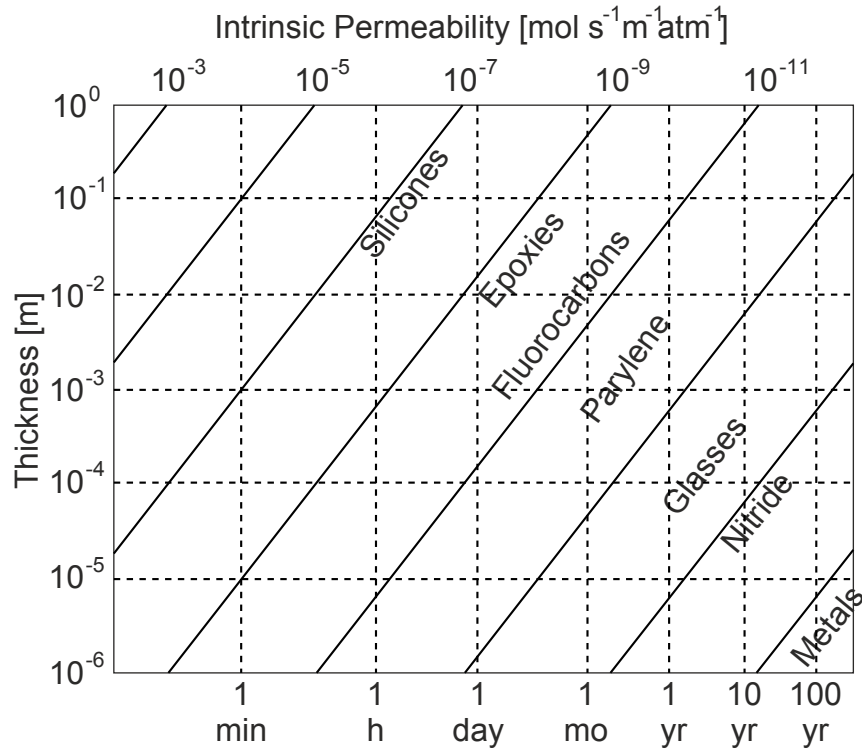


Figure 5.8: Required barrier permeability and thickness for reaching specific duration of protection (defined as the time it takes for the moisture pressure inside the package to reach 50% of the outside pressure at  $T=25^{\circ}\text{C}$  in a cavity with side  $a=1$  mm).

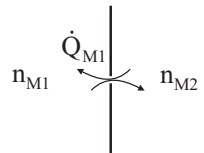
where  $N_M$  is the number of mole of gas molecule,  $n_M$  the molar density (in  $\text{mol}/\text{m}^3$ ) and  $R = 8.31\text{J}/\text{K}$ .

But what happens now if there is an aperture in the barrier? In that case, in the average, the number of molecule entering the aperture on the high density (i.e. pressure) side will be larger than on the lower density (i.e. pressure) side, resulting in a net flow of molecule from high pressure to low pressure. This flow is then simply governed by the gas molar density difference  $n_{M2} - n_{M1}$ , and can be written in the form

$$\dot{Q}_{M1} = \frac{\partial N_{M1}}{\partial t} = C(n_{M2} - n_{M1})$$

where  $C$  is expressed in  $\text{m}^3/\text{s}$  and is the conductance of the channel that takes into account the probability that a molecule is going through the aperture and not coming back.

For small hermetic package, only the study of fine leaks is of interest (larger leak will change pressure inside the package very rapidly) and as we have very narrow channels we recall the discussion in Section 3.3 and, even at standard pressure, we



may reasonably assume that  $Kn > 0.5$  and suppose that molecular flow regime is the dominant flow regime<sup>2</sup>. The channel conductance then takes a simple form as shown by Knudsen and the equation governing the flow becomes<sup>3</sup>:

$$\dot{Q}_{M1} = \frac{F_m}{\sqrt{M}}(\sqrt{T_2}n_{M2} - \sqrt{T_1}n_{M1}) \quad (5.2)$$

where we have left the opportunity for the temperature to be different on both side of the leak. In the case where the temperature is equal, the equation can be simplified further as:

$$\dot{Q}_{M1} = F_m \sqrt{\frac{T}{M}}(n_{M2} - n_{M1}) \quad (5.3)$$

where  $\dot{Q}_{M1}$  is the molar flow in region 1 counted positive if it enters the region,  $F_m$  is the molecular conductance of the leak (which for a single ideal circular channel of diameter  $d$  and length  $L$  is given by  $F_m = \sqrt{R\pi/18d^3/L}$ ),  $T$  the absolute temperature equal on both side,  $M$  the molecular mass of the gas and  $n_{Mi}$  the molar density in region  $i$ . The conductance of the conduit (also called the standard or true leak rate), for molecular flow is then defined as:

$$C = F_m \sqrt{\frac{T}{M}}$$

We note that the conductance depends on the temperature (and that this definition assumes equal temperature on both side of the leak), but if it is known for one gas, it can be obtained for any other gases, provided its molecular mass is known. Helium gas, having the smallest molecular mass after hydrogen, will leak faster than most other gases and, at the same temperature, the ratio of the conductance is given by  $C/C_{\text{He}} = \sqrt{M_{\text{He}}/M}$ . Table 5.4 gives this ratio for some commonly encountered gases. The air value is for rough computation as the molar density (or partial pressure) of each component of air should be used with their corresponding leak rate to estimate the effect of the leak (78% N<sub>2</sub>, 31% O<sub>2</sub>...). Actually, because of the difference in gas conductance, the leaked 'air' will have a different composition than normal air (N<sub>2</sub> will leak faster than O<sub>2</sub>) until the molar density equilibrates<sup>4</sup>.

The molar leak rate  $\dot{Q}_M$  is easily computed if the molar density difference is constant, but actually for a closed package a gas leaking inside will gradually increase the molar density (or lower if it leaks outside) continuously changing the leak rate, until the molar density on both side becomes equal (or more exactly the product  $n_M\sqrt{T}$ , when the temperature is different on both side).

<sup>2</sup>Diffusion flow will be dominant over molecular flow in the case where  $L \ll d$ , that is a conduit much shorter than wider between the two regions which is not a common case in encapsulation.

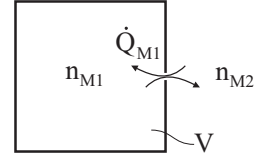
<sup>3</sup>We use here molar density instead of pressure as it makes the theory more sound and removes some ambiguities in existing derivation.

<sup>4</sup>This is actually the principle behind gas chromatography, where gas species diffuse at a speed depending on their molecular weight

Gas	Conductance ratio $C_{\text{gas}}/C_{\text{He}}$
H <sub>2</sub>	1.414
He	1
N <sub>2</sub>	0.3779
O <sub>2</sub>	0.3536
H <sub>2</sub> O	0.4714
Air	0.3804

Table 5.4: Conductance ratio for different gases with respect to He.

If we consider the standard situation for a package where the molar density (and pressure) of the gas in the environment is unaffected by what can leak from the package (i.e.,  $n_{M2} = n_{M20}$  is constant), and by using the volume  $V$  of the package to relate the mole number to molar density  $n_{M1} = N_{M1}/V$ , we can solve the flow equation Eq. 5.3 and obtain the evolution with time of the difference in molar density inside and outside the package:



$$n_{M2} - n_{M1} = n_{M20} - n_{M1} = (n_{M20} - n_{M10})e^{-\frac{C}{V}t}$$

Then the molar leak rate is obtained as:

$$\dot{Q}_{M1} = C(n_{M2} - n_{M1}) = C(n_{M20} - n_{M10})e^{-\frac{C}{V}t}$$

and the molar density inside the package evolves following:

$$n_{M1} = n_{M20} + (n_{M10} - n_{M20})e^{-\frac{C}{V}t}$$

The molar density evolution equation can be converted to use the partial pressure in the package by applying Eq. 5.1 ( $n_i = p_i/RT_i$ ), giving:

$$p_1 = p_20 + (p_10 - p_20)e^{-\frac{C}{V}t}$$

This last equation is the one generally obtained by following the standard derivation of the molecular flow theory, however it is only valid if the temperature inside and outside the package is the same (which is not necessarily the case as heat is usually generated inside the package). Actually, we see above in Eq. 5.2 that if the temperature is different, the variable of interest becomes  $n_M\sqrt{T}$  and the pressure can not be simply used anymore.

Typically, the equation is used with two different initial conditions:

- the package is in vacuum at  $t = 0$  and air slowly leak inside. Then we have  
:  $p_{\text{pack}} = p_{\text{air}}(1 - e^{-\frac{C}{V}t})$

- the package is pressurized with air at pressure  $p_0$  at  $t = 0$  and leaks outside. We have :  $p_{\text{pack}} = p_{\text{air}} + (p_0 - p_{\text{air}})e^{-\frac{C}{V}t}$

Of course, other situation will involve more complex behavior, such as, for example, when He is used to pressurize the chip: as He will escape through the leak, air will try to enter the package (the partial pressure of nitrogen or oxygen is 0 inside) - but will be impeded in its inward flow by the out-flowing He, modifying the results seen above.

---

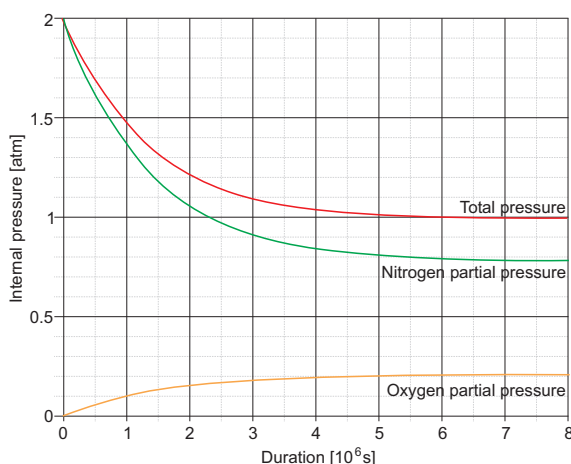
**Example 5.3** Variation of pressure in a leaking pressurized package placed in air.

**A** RF SWITCH packaged with anodic bonding has a  $500 \mu\text{m} \times 500 \mu\text{m} \times 200 \mu\text{m}$  cavity pressurized at to 2 atm with pure nitrogen ( $\text{N}_2$ ). The leak rate, measured with He, is  $10^{-16}$  mol/s (or  $2 \cdot 10^{-12}$  mbar l/s). What will be the pressure inside the package after 1 year?

We know that nitrogen will leak out but that other gases (oxygen, carbon dioxide...) will leak in. Actually, although the pressure inside the package is higher, molecular flow is driven by difference in molecular density (or partial pressure). For simplifying the problem, let's suppose we have only two gases,  $\text{N}_2$  and  $\text{O}_2$ , that do not interact together.

The leak conductance for these two gases is computed from He data using Table 5.4 :  $C_{\text{N}_2} = 0.3779 \cdot C_{\text{He}} = 0.3779 \cdot 10^{-16}$  mol/s and  $C_{\text{O}_2} = 0.3536 \cdot 10^{-16}$  mol/s.

Ambient air is composed of about 78% nitrogen and 21% oxygen, that is, the partial pressure of the gas outside the package is  $p_2^{\text{N}_2} = p_{20}^{\text{N}_2} = 0.78$  atm and  $p_2^{\text{O}_2} = p_{20}^{\text{O}_2} = 0.21$  atm, while originally inside  $p_{10}^{\text{N}_2} = 2$  atm and  $p_{10}^{\text{O}_2} = 0$  atm. This is converted to molar density by using the perfect gas law,  $n_{Mi} = p_i/RT$ , and by assuming a constant temperature of 300 K (27°C). The evolution of the molecular density inside the package for the two gases is obtained using Equation 5.2.2:



$$n_{M1}^{\text{N}_2} = n_{M20}^{\text{N}_2} + (n_{M10}^{\text{N}_2} - n_{M20}^{\text{N}_2})e^{-\frac{C_{\text{N}_2}}{V}t}$$

$$n_{M1}^{\text{O}_2} = n_{M20}^{\text{O}_2}(1 - e^{-\frac{C_{\text{O}_2}}{V}t})$$

Finally, the total pressure inside the package is the sum of the partial pressure  $p_1 = n_{M1}^{\text{N}_2} \cdot RT + n_{M1}^{\text{O}_2} \cdot RT$ , and is shown in the Figure. Clearly, after 1 year (i.e.  $31.536 \cdot 10^6$  s) the pressure, and the gas composition, inside the package will be the same as ambient air : the leak is way too big.

---

To maintain tight hermeticity the best method is probably to use wafer bonding technologies with limited permeability to gas, like glass to silicon anodic bonding or metal to silicon eutectic bonding (cf. 3.4.3).

However all MEMS can not be treated in this way, and for example the Texas Instrument DLP's packaging is more complex because the tiny mirror would not survive harsh elevated temperature treatment including glass bonding. Thus, a full chip-by-chip hermetic package in metal with a transparent glass window had to be designed. The package is sealed using brazed metal can in a clean room under a dry nitrogen atmosphere with some helium to help check leaks, and incorporates strip of getter material, a special material for removing the last trace of humidity (Fig. 5.9). The getter is a material with a high porosity that will react with the

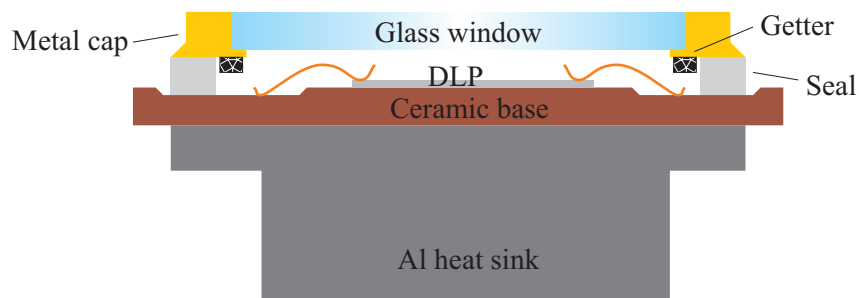


Figure 5.9: Schematic of TI's DLP package with hermetic encapsulation and getter.

target gas (usually water vapour, but oxygen or other gases can also be targeted) forming a solid compound at the getter surface. In general a special trick is used to activate the getter after the package is closed (heating it above a certain temperature, for example) so that the getter does not get quickly saturated in the open air during packaging operation. Getter can be deposited by PVD or CVD in thin films or pasted in the package, but they have a finite size and thus will work best for a finite amount of gas molecules: in general the trace of gas adsorbed on the inner surface of the package. On the longer run, they will retard the degradation due to the permeation of water vapour but even very small leaks will saturate the getter rapidly.

In view of the complexity - and cost - of fully hermetic packaging, it is lucky that all MEMS do not need such packages and the end of the 1990s Ken Gileo and others introduced a new concept: near-hermetic packaging. If this concept resists a formal definition (what leak rate defines quasi-hermeticity?) a heuristic definition would say that the package should be 'good enough' for the MEMS operation. Accordingly, relaxing the constraint on the hermeticity (possibly by supplementing it with a getter) opens up the range of techniques that can be used, and for example polymer encapsulation or bonded wafer using solder bonding or even polymer bonding, would be generally good enough while allowing a much simpler bonding procedure (as, for example, the flatness requirement with a relatively thick solder paste is heavily relaxed compared to what is needed for anodic, or even worse, for fusion bonding).

Finally, for hermetic or quasi-hermetic package the remaining question will be how to test the hermeticity? Gross leak can easily be detected by a simple bubble test where the package is heated and immersed in a liquid: if there is a gross leak the heated gas will escape through the leak and form bubble. But this test does rarely make sense in MEMS packaging, where gross leak could be spotted during visual inspection. For fine leak, the standard test is the He-leak test using what is known as the “bombig test”. Here, the package is first placed into a high pressure chamber with Helium for some time to force the gas into the package. Then the package is taken out of the high pressure chamber and the He leak rate is measured with a calibrated mass-spectrometer. This procedure is able to measure fine leak rate in package with a volume of a fraction of a  $\text{cm}^3$  down to a leak rate of about  $5 \cdot 10^{-17} \text{mol/s}$  (or  $10^{-12} \text{mbar l/s}$ ) for the best leak rate detection systems – which could be compared to the . Still, in general MEMS package are too small (and the amount of He too little) to use directly the test on real packages and special test bed (using the same material and bonding technique but having a larger cavity) need to be built to perform this test. More advanced techniques will use the measurement of the Q factor of a mechanical resonator microfabricated on the chip itself. As the pressure inside the package increases (when it is left in an ambient at normal pressure or at higher pressure for accelerated tests), the Q factor of the resonator decreases, allowing to estimate the leak rate. The advantage of this technique is that it is sensitive enough to measure the leak directly on the MEMS packages with their actual dimension.

### 5.2.3 Electrical feedthrough

The main technique used for connecting the die to the contact on the lead frame (cf. Figure 5.6), or to other dies in case of in-package assembly, is wire-bonding. Originally developed as thermocompression gold to gold bonding (still used for wafer-to-wafer bonding cf. Sec. 3.4.3), it evolved to take benefit from ultrasonic force. We can distinguish now 3 different types of techniques (Table 5.5, with the dominant one in IC manufacturing being thermosonic bonding, while ultrasonic bonding is more often used for MEMS because of its low process temperature, although the high ultrasonic energy may pose problem to mobile mechanical part.

The type of the bond (ball or wedge) depends on the tool used, as shown in Figure 5.10. Wedge bonding would usually allow higher contact density (pitch  $< 50\mu\text{m}$ ), but is slightly slower than ball bonding.

The most common material used for the bond are gold and aluminum. The wire is 50-75  $\mu\text{m}$  diameter while the pad minimum dimension should be about 4 times the wire diameter for ball bonding, or 2 times for wedge bond. The Al/Au combination may form intermetallic and develop Kirkendall voids at moderate temperature that are detrimental to the reliability of the bond. Accordingly, Al/Al or Au/Au combinations should be preferred as they are free from these problems. Other materials could be used as well in special cases. Copper wires are attractive

Technique	Pressure	Temp.	US	Mat.	Type
Thermocompression	High	300-500°C	No	Au/Au Au/Al	B-W
Ultrasonic	Low	25°C	Yes	Au/Al Au/Au Al/Al	W-W
Thermosonic	Low	100-150°C	Yes	Au/Au Au/Al	B-W W-W

Table 5.5: Comparison of wire-bonding techniques

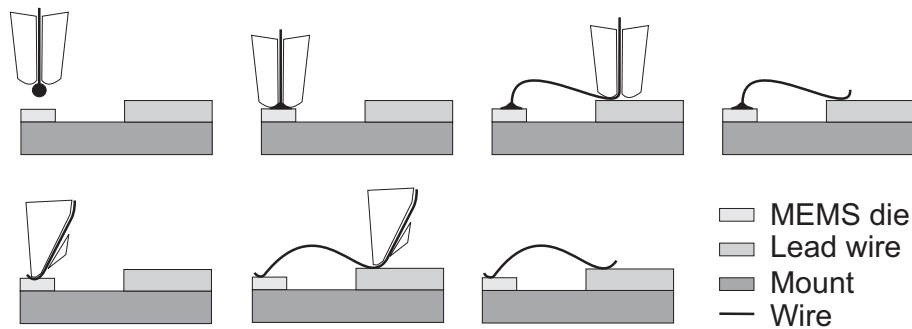


Figure 5.10: Principle of (top) Ball-Wedge bonding and (bottom) Wedge-Wedge bonding.

because of their low price and excellent conductivity, but they are harder to bond requiring more force and inert atmosphere as they oxidize readily. The Au/Ag combination has shown excellent high temperature reliability and has been used for years with Ag plated lead-frame in high temperature electronics and sensors.

Wirebonding is harder to apply if the MEMS has been capped for protection or for hermeticity. Actually, by covering the surface of the MEMS wafer, the cap wafer hides the pads that would have normally been used for the interconnect. Thus, many techniques have been developed to get the electrical contact fed under the cap. In general the techniques are split in three different classes, with typical examples in each cases shown in Figure 5.11:

**Through the cap** in that case the contact are taken through the cap. The advantage of this technique is that it does not interfere with the MEMS process, as the bulk of the additional processes needed is performed on the cap itself.

**Through the wafer** in that case the contact are fed through the MEMS wafer to its back-surface. This require additional process on the MEMS wafer, but it can be interesting to do if there are already through-the-wafer holes in the main MEMS process.



**Lateral** here the contact run below the surface of the MEMS wafer at the edge of the cap. The difficulty here is to have a conductor running on the surface that won't compromise the hermeticity of the bond. A solution proposed by Sensoror for their MPW process is to use boron doped conductors diffused at the surface of a SOI wafer before bonding, while the foundry Tronics proposes a process where epitaxy is used to bury the pass-through wiring before bonding.

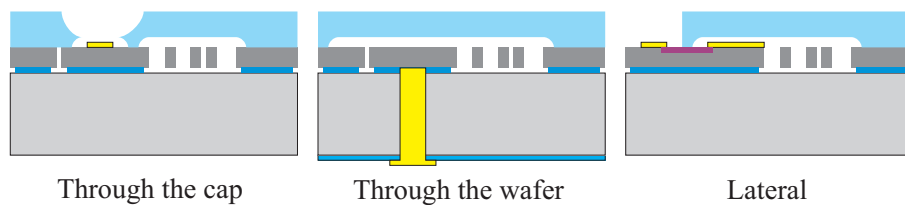


Figure 5.11: Example of electrical feedthrough techniques for capped wafers.

The wafer bonding approach, has the potential to be used for a true wafer level packaging technique (WLP) - without the need for overmolding - that could allow packaging and of all the chips on the wafer in a single batch operation. Actually the only remaining feature that is needed after wafer bonding is the possibility to connect the chip to a printed circuit board (PCB) so that it can be assembled in a complex system. This feature can be obtained by using the ball bonding technology seen in flip-chip integration that we have seen in Figure 4.2. The solder ball can be used to interconnect the IC and the MEMS chip, but also the resulting stack of chips to the PCB, a step forward in 3D packaging technique. The advantage of solder ball bonding is that it is performed in batch, as the balls will solder the two wafers after they have been simply aligned in contact with a pad and heated in an oven. Besides, as we can see in Figure 5.12, the deposition of solder balls, also called bumping, can be performed at wafer level in batch. In this case an under-bump-metalization (UBM) (for example a bi-layer of TiW/Au 150nm/300nm) is first sputtered before the solder is electroplated in a resist mold, before it can be reflowed to form spheres. The combination of ball-bonding and wafer bonding (including the inclusion of via to bring contact from one side of the wafer to the other) may result in packaging that is the same size as the die, and we then speak of chip-size packaging (CSP), the ultimate goal for extreme system miniaturization.

### 5.3 Testing and calibration

Testing is required to increase the reliability of the packaged MEMS. Different type of tests are performed during the complete process, and we distinguish qualification tests to detect failed dies while burn-in tests screen out low reliability dies. Burn-in tests are performed at chip level often after packaging, while qualification tests

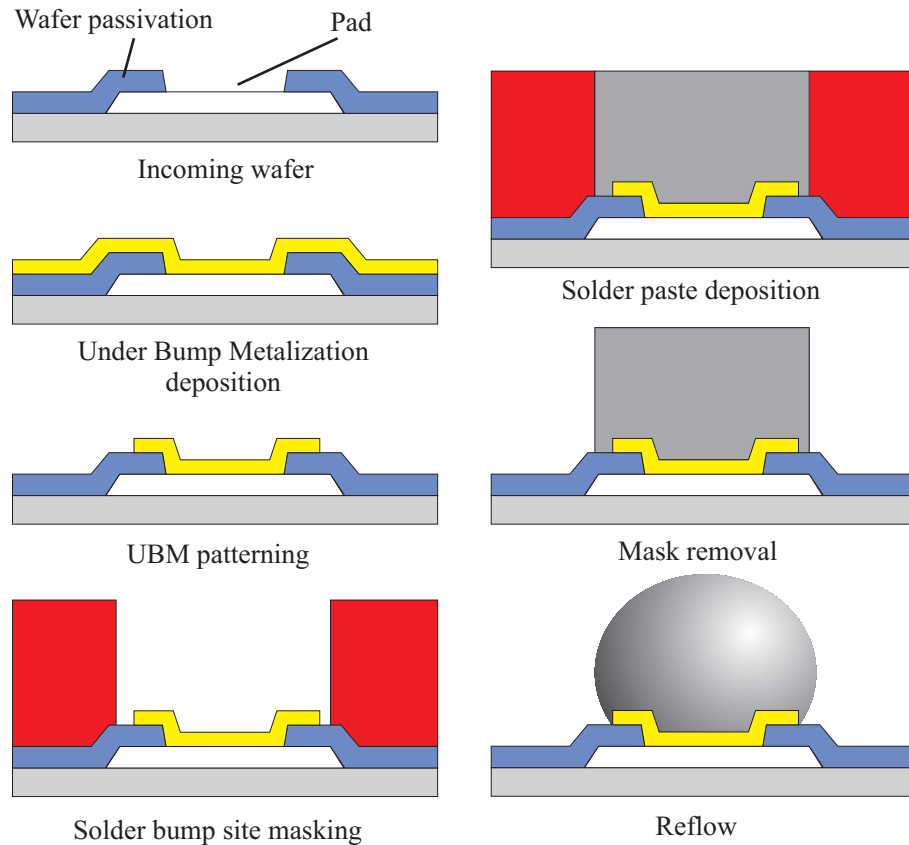


Figure 5.12: Formation of solder bump on wafer pad.

are being performed both at wafer and chip level to screen the chips that should not be packaged.

In addition to qualification test performed at wafer level, the testing phase allows to perform calibration, which will be normally conducted after packaging. The purpose of calibration is to compensate some of the defects into the MEMS characteristic to make it work more ideally. Actually the linearity of MEMS sensor or actuator may not be perfect or, more often, the cross-sensitivity with some environmental parameters (usually temperature) would require to be accounted for, another operation called compensation.

The calibration and compensation will help to decrease the influence of unavoidable defects of the micro-sensor but, most of the time, the calibration and the compensation will be decided at the factory. In the future it is expected that all the system will include a capability to perform self-calibration. If this is not always possible to implement (for example, applying a reference pressure for a pressure sensor, is not an easy task), some system have so much drift that they won't operate in any other manner. A good example is provided by some chemical micro-sensors that can only compare concentration in two fluids (thus need a calibration for each measurement) but can hardly give 'absolute' readings.

**Example 5.4** Using calibrated and compensated sensor or not?

A GOOD INSIGHT at the importance of the calibration and compensation may be gained by comparing two products from the range of pressure sensors from Motorola. The MPX10 is an uncalibrated and uncompensated pressure sensor, while the MPX2010 is passively calibrated and compensated. We report in the following table some of their characteristics, extracted from the manufacturer's datasheets (MPX10/D and MPX2010/D).

Characteristic	MPX10			MPX2010			Unit
	Min	Typ	Max	Min	Typ	Max	
Full Scale Span	20	35	50	24	25	26	mV
Offset	0	20	35	-1.0	–	1.0	mV
Temp. eff. full scale span	-18	–	-13	-1.0	–	1.0	% FS
Temp. eff. offset	–	$\pm 1.3$	–	-1.0	–	1.0	mV
Temp. coeff. full scale span	-0.22	–	-0.16				$\%/^{\circ}\text{C}$
Temp. coeff. offset	–	$\pm 15$	–				$\mu\text{V}/^{\circ}\text{C}$
Sensitivity	–	3.5	–	–	2.5	–	mV/kPa

We see the effect of calibration of the sensor using laser trimmed resistor on the full scale span, that is properly normalized, and on the offset that is almost suppressed. The compensation technique use additional resistor that are also laser trimmed to reduce tremendously the influence of temperature on the full scale span. However we could note that the effect of temperature on the offset is not really compensated this way.

What is not shown here is the large difference in price between this two sensors, which could be an important element of choice!

### 5.3.1 Testing

However if the overall reliability is mostly considered to be governed by the fabrication process, the reality is different. Actually all the process steps added during packaging and test may affect adversely the final reliability of the device in sometimes unsuspected ways. For example, it has been shown that the qualification tests performed at wafer level may affect reliability of wire bonding. Actually, these tests are performed in a probe station using sharp needles to contact the pads and apply different test signals to the device. The contact of the test probe on the gold pads would leave a small scratch on the pad surface, which has been shown to affect the bonding strength, ultimately possibly decreasing the reliability of the packaged system.

The main problem faced by MEMS testing is that we now have to handle signal that are not purely electrical, but optical, fluidic, mechanical, chemical... Then, verifying the absence of defect needs the development of specialized system and

new strategies.

Texas Instrument DLP chip may have as many as 2 millions mirrors and simple math shows that testing them one by one during 1 s would take approximately three weeks at 24 h/day – clearly not a manageable solution. TI has thus developed a series of test using specific mirror activation patterns that allow testing mirrors by group and still detect defect for individual mirror, like sticking or missing mirror. After testing at wafer level the chip is diced, put into packages and then goes through a burn-in procedure. They are then tested again before being finally approved. TI noticed that the encapsulation step decreased the yield if the environment wasn't clean enough and they have to use a class 10 clean-room for the packaging of their DLP chips.

Testing is also a major hurdle for micro-sensor and to facilitate it additional test features may have to be included into the MEMS design. A good example is given by the integrated accelerometer range from Analog Devices. The system use a micromachined suspended mass whose displacement is monitored by using induced change in capacitance. However one part of this electrodes has been configured as an *actuator*. By applying a voltage on this electrodes it is possible to induce a movement of the mass *without external acceleration*. This is used before the packaging to verify the mechanical integrity of the accelerometer... and allow to save a lot of money, compared to a set-up that would need to apply a *real* acceleration. Moreover that function may be used during operation in a smart system to verify the integrity of the accelerometer.

Testing is conducted at different stage during the fabrication of the MEMS, but the final test are normally conducted *after* the packaging is done. The reason is simple: the packaging process always introduce stress (or damping) that change the characteristics of the sensing elements, but need to be accounted for. This final test can be used for burn-in and usually allow the final calibration and compensation of the sensor.

### 5.3.2 Calibration

The calibration of the MEMS is the adjustment needed to deliver the most linear possible transfer function, where the output goes from 0% to 100% when the input varies in the same range. It generally means three different things:

1. linearization of the transfer function (i.e., having a constant sensitivity)
2. adjustment of the output span (i.e., adjusting the sensitivity)
3. suppression of the offset

It is a particularly important step for microsensors, although it is understood that all the other environmental variables (e.g., temperature, humidity, pressure, stress...) are kept constant during the calibration and will be specified. As such, the response of the system will generally be the best at calibration point. We note

that the cancellation of the change caused by these other environment variables on the sensitivity or on the offset is left to the compensation procedure. Thus it is possible to have a calibrated but uncompensated system, while the reverse is often much less interesting.

To perform these tasks it is possible to trim integrated additional analog circuit or to use a digital signal and a CPU. The choice between both is a mix of speed, complexity and cost analysis, but obviously the integrated analog approach needs more wit! Obviously it is possible to mix the two methods, that are not mutually exclusive, and for example offset removing is often performed analogically while sensitivity trimming is more easily done with digital techniques. In both case the calibration is performed after having the result of the calibration tests that will tell to which extent the sensor needs to be adjusted.

With analog calibration technique, the basic tuning elements are trimmable resistors. The technique use a laser or electric fuse to trim the value of resistor controlling MEMS sensitivity and offset. This method has the advantage to be relatively cheap and to provide sensor with very high speed. Actually the laser trimming method is less useful as it does not allow recalibration nor allow the calibration after packaging, and electrical trimming is preferred. The trimmable resistor does not allow to compensate for complex non-linearity, and if previously there was a lot of effort devoted in developing linearizing scheme with complex analog circuits (see for example Section 4.4.2), now the trend is toward digital calibration for the more complex cases.

Digital calibration technique is quite straightforward to implement if sufficient computation power is available, but will always be slower because it needs analog to digital conversion and data processing. The principle is to perform a calibration test, and to use the data recorded to compute the output of the sensor. The CPU may use a model of the sensor, usually implemented using a high order polynomial, or use the actual values of the calibration test, stored in a look-up table (LUT). Between two points of the look-up table, the correction is estimated by using a linear approximation. In principle this approach allows to compensate for any non-linearity, and should deliver ‘perfect’ characteristics. However apart from the speed problem, the analog to digital conversion of the signal introduce new parameters that does not allow to correct all the defect in the transfer function characteristic. For example, we show in Figure 5.13 the case of a MEMS sensing element presenting a flatter part in his transfer function. The marked non-linearity of this transfer function introduce measurement errors than can not be calibrated out. Actually, any input laying in the range between about 5 and 12.5, will give the same A/D converter output: [0110]! We have here an important loss in the accuracy of the sensor that can not be suppressed. Thus if digital calibration allows to correct many defects, it is not possible to eliminate all of them. From this example, we may remember that the maximum error presented by non-linear element will be governed by the region of the transfer function presenting the smallest slope.

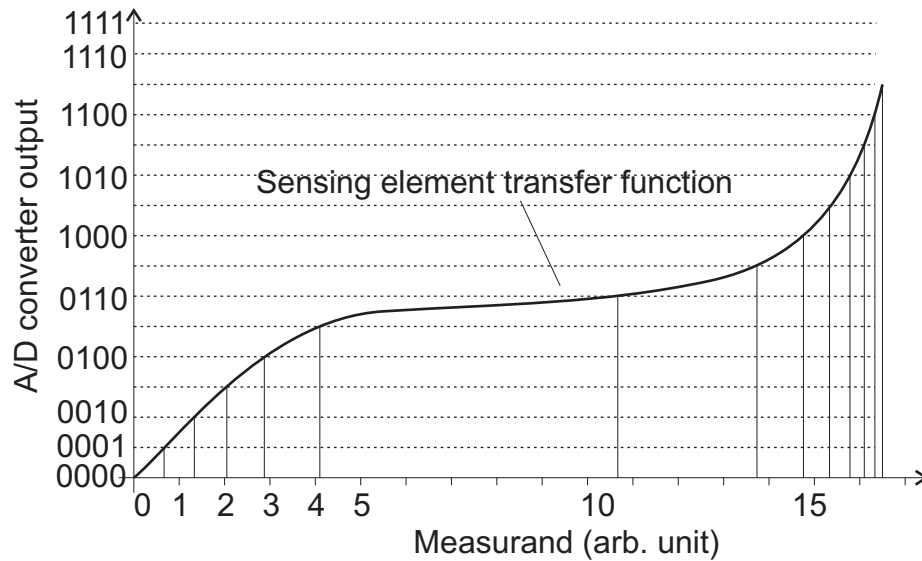


Figure 5.13: Effect of sensing element characteristic on accuracy after digital conversion

Additionally, the digital converter works best by using its full range (otherwise, we ‘loose’ some bits of resolution)... which is not necessarily the same range as the MEMS operation range. In general, it is thus necessary to have a controllable gain amplifier that will adapt the converter signal to the MEMS range. Additionally it would be interesting for the same reason to have a controllable way to trim an offset voltage. These two functions can be merged together in an amplifier with tunable gain and offset. As an example, Microsensors Inc. is producing an integrated circuit for sensor interfacing, the MS3110, that has an output amplifier with a voltage gain that may be chosen between 1.7 and 2.3 using 8 bits of control for a step of 0.0024 and a trimmable offset that change in a range of  $\pm 100$  mV using 5 bits of control with 6.25 mV step. Actually, this component will also be very useful for the compensation of the sensor, as we will see now.

### 5.3.3 Compensation

The compensation refers to the techniques used to separate the MEMS output from an interfering environmental parameter, like the temperature, and we distinguish:

**structural compensation** where preventive measure are taken at the design stage to decrease the magnitude of cross-sensitivities;

**monitored compensation** where implicit or explicit measurement of the interfering parameter allow to modify the output of the sensor to compensate for its effect. Implicit measurement approach use additional integrated circuitry with the sensor, while explicit measurement use a sensor for the interfering

parameter (e.g., a temperature sensor) and a CPU to adjust the raw measurements according to the value of the interfering parameter.

In the first class of compensation, the layout of the microsystem is important and for example it is possible to insulate it thermally to decrease the influence of temperature. Another very simple structural compensation technique that should not be underestimated is the use of symmetry in the design. The principle is to use a difference signal, while all the other interfering variables will produce a common mode signal, that will thus not be apparent on the MEMS output. The compensation for residual stress in many MEMS sensors is often based on this principle, and observing the design of the sensitive elements will invariably show a marked 2-folds or 4-folds symmetry.

The monitored compensation may be performed at the system level without explicit measurement of the perturbing parameter using completely analog signal, and we talk of implicit compensation, or with explicit compensation. In this latter case the compensation is using a CPU and an amplifier with programmable gain and offset. The choice between these two approaches is often dictated by the complexity of implementation. An implicit compensation is often used to compensate for temperature in pressure sensor, but as soon as the dependence on the external factor is complex, or when a very precise compensation is needed the compensation is performed at the system level using a CPU. Generally speaking, an implicit compensation needs more cleverness than an explicit approach that will work ‘all the time’. For example, if it is relatively easy to compensate for an offset induced by the temperature with analog circuit, a change in the gain of the transfer function of the sensor will be much more difficult to compensate and explicit compensation with analog or digital techniques will be required. Still implicit approach will produce smaller control circuit that could be a decisive advantage.

For example, implicit monitored compensation can be applied to shelter from fluctuation in power voltage. In this case the idea is simply to use working principles that are ratiometric to this voltage.

For example, a Wheatstone bridge (see Sec. 4.4.1) delivers a voltage that is proportional to the supply voltage  $V_{in}$ :

$$V_{out} \approx \frac{V_{in}}{4R} \Delta R.$$

This may seem to be a serious problem as any fluctuation in the supply voltage will result in loss of accuracy, but when we consider the complete measurement chain, this may turn to an advantage. If we consider a pressure sensor based on piezoresistive elements in a Wheatstone bridge, the signal from the sensor needs digitization further down in the measurement chain to be transmitted or recorded. The digitization based on A/D converter requires generally a reference voltage from which the quantization step (quantum) is derived. If for this reference voltage we use the bridge supply voltage  $V_{in}$ , then any fluctuation in this voltage is automatically compensated by an equivalent change in the A/D converter quantum. In

this way the recorded digital information of the pressure will be accurate even if the supply voltage change due to battery charge dwindling or other environmental factors.

A typical example of a circuit using explicit compensation (also called digital compensation) is shown in Figure 5.14. Here the CPU constantly change the

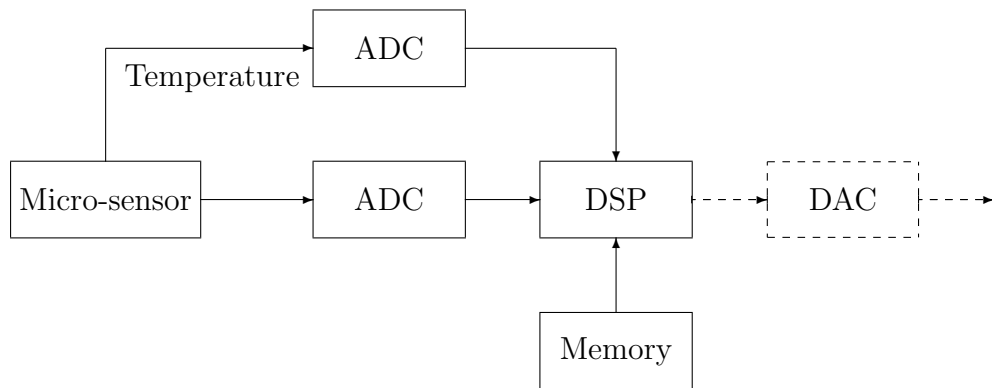


Figure 5.14: Explicit temperature compensation in a Smart Sensor

output according to the value of the input, using a model or a calibration curve. However if this implementation looks simple it has some marked drawbacks, when, for example, the sensitivity of the system decrease with the interfering parameter as shown for a sensor in figure 5.15. If the sensitivity increase ( $T_1 > T$ ) in this

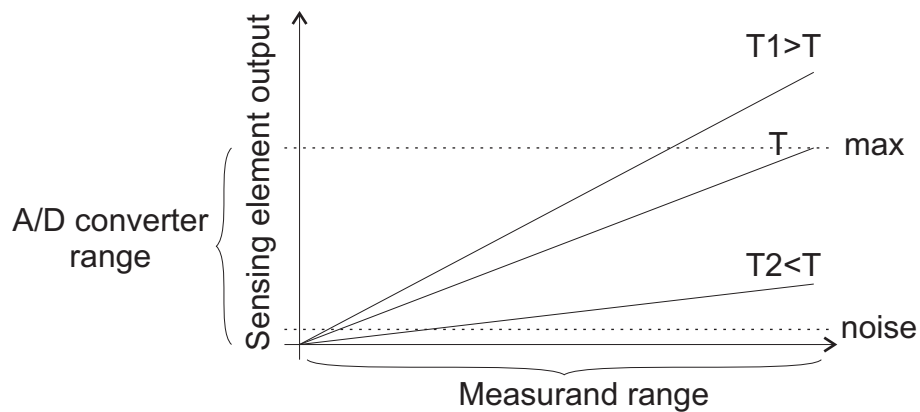


Figure 5.15: Effect of temperature on a sensor sensitivity.

simple scheme, a loss of resolution occurs because the ADC has a limited span and thus saturates. If the sensitivity decreases ( $T_2 < T$ ), and the noise at the output of the sensing element remains constant, the resolution of the sensor drops tremendously. Thus this 'all digital' approach is not completely satisfactory and a mixed digital-analog approach is required. It generally uses a tunable amplifier



or voltage source to compensate the output of the sensing element, performing also the signal standardization to match the ADC. A typical implementation of this strategy will use a digital to analog converter to control the supply voltage of the sensing element, controlling effectively its sensitivity, and use a gain tunable amplifier (GTA). The additional circuitry used is based on a digital to analog converter (DAC), converting the digital signal from the CPU to an analog signal that is used directly (voltage offset control) or use an additional compensation actuator to convert it to a usable signal (voltage/gain conversion, voltage/current conversion, voltage/force conversion as in the ADXL105). Additionally in such circuit using explicit compensation, in order to reduce circuitry, the temperature sensor and the sensing element are sharing the same A/D converter using a multiplexer.

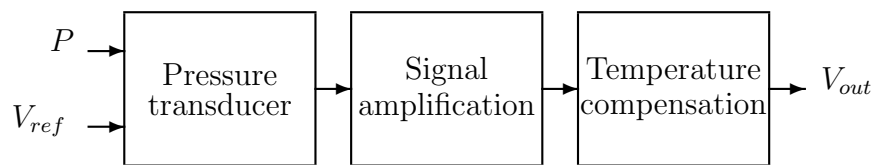
The digital signal processor (DSP) will perform the necessary computation to compute the value of the different adjustable parameters, either using a physical or empirical model of the dependence of the sensitivity and offset with the temperature, or, better, using the result from a calibration test. In this latter case, as it is not possible to perform a calibration for *every* temperature, an interpolation will be used to determine the change between two measured points. As in the case of the calibration the correction factor will generally be stored as a series of values, but in this case in a two dimensional array, the first dimension corresponding to the measurand and the second, the temperature.

The LUT approach is interesting when only one interfering parameter needs to be compensated, but when there are multiple interfering parameters, as in the case of chemical sensors, the approach becomes quickly intractable. Actually, as it is not possible to perform all the calibration tests needed, the compensation can only be done using a model. This model will be implemented in the system using a series of coefficients of a polynomial (normally with order  $< 7$ ), limiting the necessary amount of memory at the cost of a larger computational effort.

**Example 5.5** Delphi Automotive Fuel Vapor Pressure Sensor (FVPS) packaging.

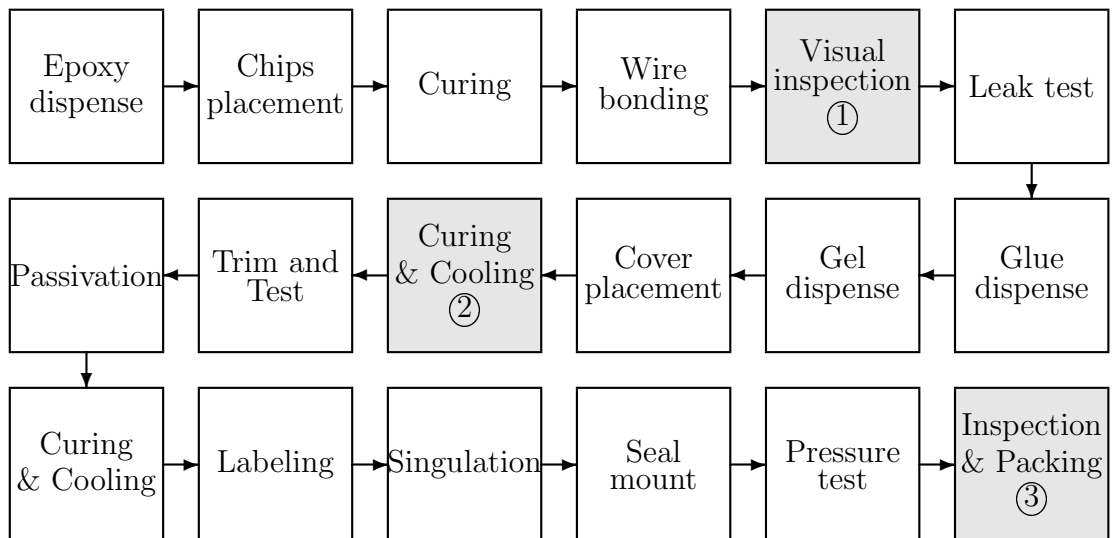
WE WILL describe the packaging operation for a low pressure sensor from Delphi<sup>®</sup> Automotive, the Fuel Vapour Pressure Senso (FVPS).

The sensor is used to check for petrol vapour leakage in car tank and provides an analog voltage output proportional to the pressure. It is based on piezoresistive sensing within a Wheatstone bridge configuration of the deformation of a thin silicon micromachined membrane. The Wheatstone bridge output is ratiometric to a reference voltage  $V_{ref}$  and proportional to the pressure  $P$ . This smart sensor also integrates a circuit for amplifying this signal to the range and offset required by the customer and for providing temperature compensation.



The design of the sensor relies on hybrid integration in the package (System In the Package - SIP). It consists in a MEMS chip (with the membrane and integrated piezoresistors) and an IC chip (a proprietary ASIC) for signal amplification and temperature compensation.

After the overmoulding step, where the polyester package is molded over the leadframe, the packaging of the FVPS would more or less have 18 different steps:



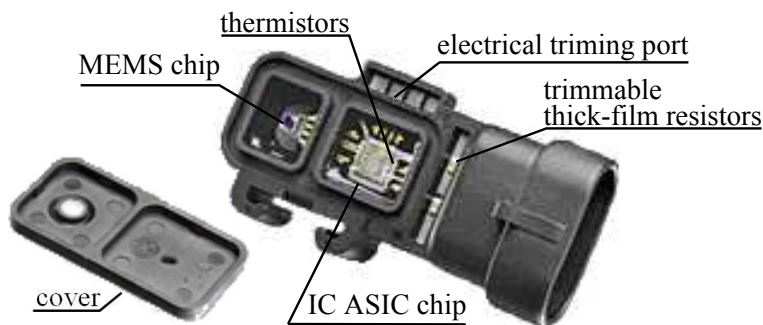
The chips are first fixed to the leadframe using adhesive. Particular care has to be taken for the MEMS chip as stress arising from thermal expansion mismatch between the polyester housing and the silicon chip or from external stress applied to the housing could affect the pressure reading. Accordingly soft adhesive are used and will take up most of the dimensional change.

---

**Example 5.5** Delphi Automotive Fuel Vapor Pressure Sensor (FVPS) packaging. (continued)

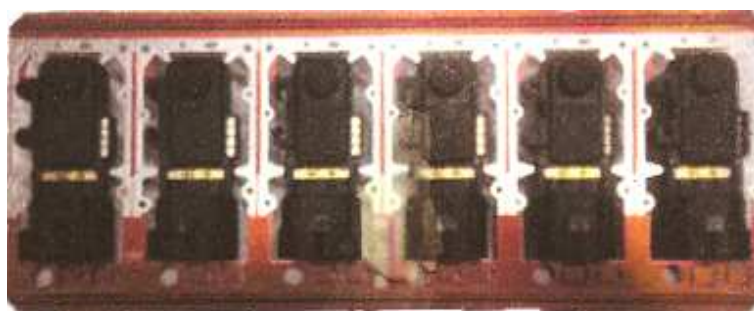
---

Then a leak test is performed to make certain the MEMS chip has been properly attached to the pressure inlet. We notice that the package chosen provides ample space around the MEMS and the ASIC chips, nicely decoupling the pressure transducer from the mechanical stress produced when the sensor is mounted in the car. The chips are then electrically connected using wire bonding to the leadframe (step ① and next figure).



Some gel is then placed to protect the sensor surface and then the cover is put in place hiding the chips while the two trimming ports are left exposed. At this step the engineers choose to allow for modularity of the pressure sensor packaging. Actually the pressure sensor may be configured in gauge (atmospheric) or differential pressure sensing. In this later configuration the cover (shown in the Figure above in front of the sensor) would have, instead of the simple vent shown here for a gauge sensor, a pressure access port, similar to what can be seen on the sensor front.

When the sensor arrives for trimming and testing it is still mounted on a lead frame in series of 6 sensors (step ② and figure below) .



This trimming step is needed for obtaining a precise and accurate sensor even with the unavoidable random changes in fabrication process. The calibration of the sensor helps providing an output as per customer specification with fixed sensitivity (or output range) and offset in the chosen pressure range.

---

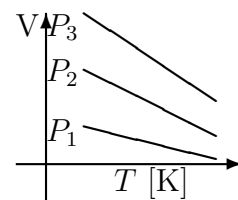
---

**Example 5.5** Delphi Automotive Fuel Vapor Pressure Sensor (FVPS) packaging. (continued)
 

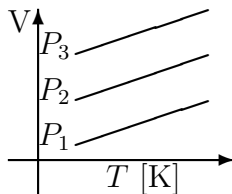
---

For the FVPS the full scale gauge pressure (i.e., with respect to ambient pressure) range is between -3.5 kPa and 1.5 kPa. Typically, for an ‘analog TTL’ output, the calibration will set the gain for an output range of 5 V with an offset of 2.5 V, resulting in a full scale output between 0 V and 5 V. Moreover, temperature is known to have a strong effect on piezoresistor as increased temperature decreases the mobility of the carrier in the doped semi-conductor. This changes the absolute value of the resistance at a fixed pressure and additionally lowers the stress/resistance sensitivity.

For achieving the correct full scale output and temperature-independence the sensor requires additional smart circuits. Actually there are dozen of different schemes for calibration and temperature compensation of piezoresistive bridge, either with analog circuits, using passive or active network, or with digital circuits. From what can be gathered from gray literature and by reverse engineering, Delphi choose in this older sensor design to follow an active analog approach, based on the use of thermistor and laser trimming of thick film resistor.



The raw output of the sensor as a function of pressure and temperature is actually showing sensitivity and offset dependance with temperature as shown in the inset.



First, the variation with the pressure of the response slope as a function of the temperature is canceled by adjusting the bridge excitation voltage. Actually the  $V_{ref}$  voltage is passively adjusted through a voltage divider made with two carefully chosen thermistors exhibiting positive temperature coefficient that are placed in series between the fixed power supply ports and the excitation ports of the bridge. Identical

network of thermistor are used on both excitation ports of the bridge so that the common mode voltage of the sensor signal remains constant over temperature. The two thermistors can be seen on the right of the ASIC chip in the figure above before the cover is placed and we observe that the metallic leadframe guaranty a good thermal path between all these elements that should be kept at the same temperature. After this operation the sensitivity of the transducer with temperature becomes independent of the pressure as shown in the inset. In a second step the signal from the bridge need to be calibrated for providing the required full scale output and also still require a final step toward full temperature compensation. Actually this is realized by adjusting the resistor network around the operational amplifier circuit placed in the package.

---

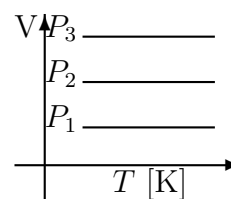
---

**Example 5.5** Delphi Automotive Fuel Vapor Pressure Sensor (FVPS) packaging. (end)

---

The procedure starts on a specially dedicated trimming station that allows applying different pressures (here a repetition of 3 measurements at -1.3 kPa and 0 Pa) and temperatures (here room temperature and in an oven at 60°C) to the sensor. The characteristics of the circuit are recorded through a hidden port connected to the ASIC electronic chip, that give access to the evolution of internal and output voltage with the pressure and temperature change. Moreover the station allows measuring the nominal value of two thick film resistors deposited on the lead frame next to the ASIC circuit. Each pressure measurements takes less than 2 s. The temperature stabilization in the oven takes the longest time and it is certainly one of the reason why 60°C has been chosen, instead of 80°C that could have been expected for a complete automotive specification: the temperature swing is smaller and takes less time. From this series of measurements, adjustment are decided and the two thick-film resistors are adjusted for obtaining the expected correction.

Pioneered in the mid 1980's – a more modern approach would use electrically trimmed resistor – the laser trimming method works by modifying the cross-section of a thin-film or thick-film resistor, and hence its resistance, by ablation with a YAG laser. The procedure allows control of resistance value within 1% or below, but it can only increase the value as the laser cuts through the material, always decreasing the



conductor cross-section. Note that for having less dependence with temperature, the trimmable resistors are normally kept insensitive to temperature changes and are manufactured with special material (Ni-Cr, Ta-Ni for thin-films and fired ceramic with metal-oxide particle like RuO<sub>2</sub> for thick-films) that can yield a TCR of 0.01/0.005%/K. The procedure ends up with a temperature-compensated sensor output (shown in the inset) fully calibrated to the clients need.

Altogether the measurements and correction must take less than 1 min in the fully automated trimming station that works simultaneously in two lanes on 6 different sensors : that is at best  $24 \times 60 \times 6 = 8640$  sensors per day for one station – and in practice much less as we need to include the other – mostly serial – tests in the packaging procedure.

After this important step, the trimming input port are permanently sealed. It is important that this last sealing and curing of polymer does not affect the transducer – or the costly calibration step would be made ineffective – and we notice in the layout that the trimming ports are far from the stress sensitive chip.

The sensor is then separated from its neighbors on the lead-frame, endures its final leak and electrical tests before it becomes ready for shipping.

---

## Problems

1. Redo the problem in Example 5.3 considering now that the electronic circuit inside the package raises the temperature to 50°C during operation, while the external temperature is considered to be at 25°C. You will need to forget about the pressure variable and derive solutions from Eq. 5.2 using  $n_M\sqrt{T}$  as the variable.

# Chapter 6

## Challenges, trends, and conclusions

### 6.1 MEMS current challenges

Although some products like pressure sensors have been produced since the 1980s, MEMS industry is, in many ways, still a young industry. The heavily segmented market is probably the main reason why a consortium like SEMI is still to appear for MEMS. However everybody agrees that better cooperation and planning has to happen if the cost of the assembly, test and packaging is to come down. MEMS can currently only look with envy as IC industry seriously considers producing RFID chips for cents - including packaging.

Again the path shown by the IC industry can serve as a model, and standardization to insure packaging compatibility between different MEMS chip manufacturers seems the way to go. Considering the smaller market size of most MEMS component, standard is the only way to bring the numbers where unit packaging price is reduced substantially. This implies of course automating assembly by defining standard chip handling procedure, and probably standard testing procedure.

Of course, the diversity of MEMS market makes it impracticable to develop a one-fit-all packaging solution and the division in a few classes (inertia, gas, fluidic) is to be expected. For example, several proposals for a generic solution to fluidic interfacing have been proposed and could become a recommendation in the future. In the other hand it is not clear if standardization of MEMS fabrication process à la CMOS will ever happen - and is even possible. But currently most of the cost for MEMS component happens during back-end process, thus it is by standardizing interfaces that most savings can be expected.

The relatively long development cycle for a MEMS component is also a hurdle that needs to be lowered if we want more company to embrace the technology.

One answer lies with the MEMS designing tool providers. The possibility to do software verification up to the component level would certainly be a breakthrough that is now only possible for a limited set of cases.

But it is also true that the answer to proper design is not solely in the hand of better computer software but also in better training of the design engineer. In particular we hope that this short introduction has shown that specific training is needed for MEMS engineers, where knowledge of mechanical and material engineering supplements electronic engineering. Actually, experience has often revealed that an electronic engineer with no understanding of physical aspect of MEMS is a mean MEMS designer.

## 6.2 Future trends in MEMS

Looking in the crystal ball for MEMS market has shown to be a deceptive work, but current emerging tendencies may help foresee what will happen in the medium term.

From the manufacturer point of view, a quest for lowering manufacturing cost will hopefully result in standardization of the MEMS interfacing as we discussed earlier, but finally will lead to pursue less expensive micro-fabrication method than photolithography. Different flavors of soft-lithography are solid contenders here and micro-fluidic and BioMEMS are already starting to experience this change. Another possibility for reducing cost will be integration with electronics - but, as we already discussed, the system-on-a-chip approach may not be optimal in many cases. Still, one likely good candidate for integration will be the fabrication of a single-chip wireless communication system, using MEMS switch and surface high-Q component.

From the market side, MEMS will undoubtedly invade more and more consumer products. The recent use of accelerometer in cameras, handphone or in the Segway is a clear demonstration of the larger applicability of the MEMS solutions - and as the prices drop, this trend should increase in the future. Of course medical application can be expected to be a major driver too, but here the stringent requirements make the progress slow. In the mid-term, before micromachines can wade in the human body to repair or measure, biomedical sensors to be used by doctors or, more interesting, by patients are expected to become an important market.

A farthest opportunity for MEMS lies probably in nanotechnology. Actually, nanotechnology is bringing a lot of hope - and some hype - but current fabrication techniques are definitely not ready for production. MEMS will play a role by interfacing nano-scale with meso-scale systems, and by providing tools to produce nano-patterns at an affordable price.

## 6.3 Conclusion

The MEMS industry thought it had found the killer application when at the turn of the millennium 10's of startups rushed to join the fiber telecommunication



bandwagon. Alas, the burst of the telecommunication bubble has reminded people that in business it is not enough to have a product to be successful - you need customers.

Now the industry has set more modest goals, and if the pace of development is no more exponential it remains solid at 2 digits, with MEMS constantly invading more and more markets. Although the MEMS business with an intrinsically segmented structure will most probably never see the emergence of an Intel we can be sure that the future for MEMS is bright. At least, as R. Feynman[30] stated boldly in his famous 1959 talk which inspired some of the MEMS pioneers, because, indeed, "there's plenty of room at the bottom"!



# Appendix A

## Readings and References

### A.1 Conferences

**MEMS conference** THE annual MEMS conference, single session and top research, happening usually in January with deadline in August. Hard to get a paper accepted, but worth it.

**Transducer's conference** The biennial conference, huge, multisession, happening every two years around June with deadline in Winter. Shows a really nice panorama of all the research in MEMS and in related fields.

**MicroTAS conference** The annual BioMEMS/Microfluidics conference with top results and researchers.

**Optical MEMS & Nanophotonics conference** The annual IEEE MOEMS conference including nanophotonics session.

**PowerMEMS conference** This annual conference is all about power generation, dissipation, harvesting, and thermal management.

**Micro-Mechanics Europe (MME) conference** The European conference on micro has a unique format for fostering interaction between participants: there are no oral formal presentation but only short introduction and discussion around posters.

**HARMST conference** The biennial High Aspect Ratio Microstructure Technology conference is rightly focused on process.

**Euroensors conference** The annual European microsensors conference.

**Micro-Nano-Engineering (MNE) conference** A European conference with a good mix of topics related to MEMS.

## A.2 Online resources and journals

<http://www.smalltimes.com/> the free international press organ of the MEMS–NEMS community.

<http://www.mstnews.de/> free news journal on European microsystem technology

<http://www.dbanks.demon.co.uk/ueng/> D. Banks renowned "Introduction to microengineering" website with plenty of information.

<http://www.aero.org/publications/aeropress/Helvajian/> The first chapter of 'Microengineering Aerospace Systems' co-authored by M. Mehregany and S. Roy and edited by H. Helvajian is online and makes a short, although slightly outdated, introduction to MEMS.

**IEEE/ASME Journal of MEMS** This journal originally edited by W. Trimmer, is arguably one of the best journal in the field of MEMS (<http://www.ieee.org/organizations/pubs/transactions/jms.htm>).

**Sensors and Actuators A** That is the most cited journal in the field, with a copious variety of research work ([http://www.elsevier.com/wps/product/cws\\_home/504103](http://www.elsevier.com/wps/product/cws_home/504103)).

**Sensors and Actuators B** Catering mostly for Chemical Sensor papers, they also have issue on MicroTAS, where you find numerous microfluidics and Bio-MEMS papers ([http://www.elsevier.com/wps/product/cws\\_home/504104](http://www.elsevier.com/wps/product/cws_home/504104)).

**Micromachines** Arguably the best *open* journal on the MEMS topic edited by MDPI - definitely a must for those who care about knowledge freedom, like we do at Memscyclopedia.org. Note that its current editor in chief (2018) is N.-T. Nguyen, a highly cited professor in microfluidics related topics (<http://www.mdpi.com/journal/micromachines>).

**Journal of Micromechanics and Microengineering** Another top journal, edited in Europe by IOP with all types of MEMS (<http://www.iop.org/EJ/S/3/176/journal/0960-1317>).

**Smart Materials and Structures** Another IOP journal, with editor V. Varadan, that has a more material oriented approach than his cousin (<http://www.iop.org/EJ/S/3/176/journal/0964-1726>).

**Microsystem Technologies** A Springer journal that favors papers on fabrication technology and particularly on high-aspect ratio (LIGA like) technology (<http://link.springer.de/link/service/journals/00542/index.htm>).

- Journal of Microlithography, Microfabrication, and Microsystems Journal** from the SPIE (<http://www.spie.org/app/Publications/index.cfm?fuseaction=journals&type=jm3>).
- Sensor Letters** A journal covering all aspects of (micro)sensors science and technology (<http://www.aspbs.com/sensorlett/>).
- Lab on a chip** The top bioMEMS/microfluidics journal (<http://pubs.rsc.org/en/journals/journalissues/lc>).
- Biomicrofluidics** An AIP journal that has a good mix between microfluidics and biological applications (<https://aip.scitation.org/journal/bmf>).
- Microfluidics and Nanofluidics** One of the top journal for microfluidics and related technologies (<http://www.springerlink.com/content/1613-4982/>).
- Experimental Thermal and Fluid Science** A good journal geared toward more fundamental issues in microfluidics (<http://www.journals.elsevier.com/experimental-thermal-and-fluid-science/>).
- Biomedical microdevices** A Bio-MEMS journal (<http://www.wkap.nl/journalhome.htm/1387-2176>).
- Biosensors and Bioelectronics** Another Bio-MEMS journal with more emphasis on sensors ([http://www.elsevier.com/wps/product/cws\\_home/405913](http://www.elsevier.com/wps/product/cws_home/405913)).
- IEEE Transaction on Biomedical engineering** Bio-MEMS and biomedical application can be found here (<http://www.ieee.org/organizations/pubs/transactions/tbe.htm>).
- IEEE Photonics Technology Letters** Highly cited photonics journal publishing short papers, including optical MEMS or MOEMS (<http://www.ieee.org/organizations/pubs/transactions/ptl.htm>).
- IEEE/OSA Journal of Lightwave Technology** A good quality photonics journal regularly featuring some optical MEMS work (<http://www.ieee.org/organizations/pubs/transactions/jlt.htm>).

### A.3 Other MEMS ressources

<http://www.memsnet.org/> the MEMS and nanotechnology clearinghouse.

<http://www.memsindustrygroup.org/> the MEMS industry group aimed at becoming a unifying resource for the MEMS industry. Will hopefully initiate standardization effort and eventually establish a MEMS roadmap.

<http://www.yole.fr/> one of the MEMS industry watch group publishing regular report on the market.



# Appendix B

## Causality in linear systems

In linear system the relationship between the input  $x$  and the output  $y$  can be represented by a differential equation:

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \dots + a_1 \frac{dy}{dt} + a_0 y = b_m \frac{d^m x}{dt^m} + b_{m-1} \frac{d^{m-1} x}{dt^{m-1}} + \dots + b_1 \frac{dx}{dt} + b_0 x \quad (\text{B.1})$$

where the sum  $m+n$  is called the order of the system with the important condition that normally  $n > m$ . This condition describes the assumption of causality in the system: the output is created by the input, not the reverse!

This may be better understood if instead of looking at the derivative, we ‘invert’ the problem and use integration. Let’s take for example  $n = 3 > m = 2$  :

$$a_3 \frac{d^3 y}{dt^3} + a_2 \frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = b_2 \frac{d^2 x}{dt^2} + b_1 \frac{dx}{dt} + b_0 x$$

integrate three times and reorganize the equation.

$$\begin{aligned} a_3 y + a_2 \int y dt + a_1 \iint y dt + a_0 \iiint y dt &= b_2 \int x dt + b_1 \iint x dt + b_0 \iiint x dt \\ y &= \frac{b_2}{a_3} \int x dt + \frac{b_1}{a_3} \iint x dt + \frac{b_0}{a_3} \iiint x dt - \frac{a_2}{a_3} \int y dt - \frac{a_1}{a_3} \iint y dt - \frac{a_0}{a_3} \iiint y dt \end{aligned}$$

Thus the *output*  $y$  becomes a linear combination (i.e., a weighted sum) of the integral of the input and the output itself. Integration is a causal operator (i.e., a sum over time starting at  $t = 0$ ) and the function can be physically implemented. Now if we exchange the values of  $m$  and  $n$ , that is  $n = 2 < m = 3$  we have:

$$\begin{aligned} a_2 \frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y &= b_3 \frac{d^3 x}{dt^3} + b_2 \frac{d^2 x}{dt^2} + b_1 \frac{dx}{dt} + b_0 x \\ a_2 \int y dt + a_1 \iint y dt + a_0 \iiint y dt &= b_3 x + b_2 \int x dt + b_1 \iint x dt + b_0 \iiint x dt \\ x &= \frac{b_2}{b_3} \int x dt + \frac{b_1}{b_3} \iint x dt + \frac{b_0}{b_3} \iiint x dt - \frac{a_2}{b_3} \int y dt - \frac{a_1}{b_3} \iint y dt - \frac{a_0}{b_3} \iiint y dt \end{aligned}$$

thus after 3 integrations we find that the *input*  $x$  becomes a linear combination including integral of the output, and thus the *input* would depend on what happened previously at the *output*! Certainly not a causal behavior...

If  $n = m$  we will find that the output is directly proportional to the input, implying instantaneous transmission of information through the system, which should violate Einstein postulate. However, the systems we study using block or circuit analysis are punctual systems: they have no physical size, thus signal *can* travel instantaneously from the input to the output. We can more easily see that we use this simplification all the time by considering a simple voltage divider. In this circuit the relationship between input and output is given by  $y = \frac{R_1}{R_1 + R_2}x$ , meaning that as soon as the input voltage change the output will change, which is clearly 'not physical'.

In practice, we will often use systems where  $m = n$  as in the case of Example 2.2, but if we are rigorous, such systems can not exist, and a delay should be introduced to model the relationship between the output and the input of a system.

Finally it should be noted that this physical delay is of a different nature (i.e. speed of propagation of the information) that what we observe in RC circuit (i.e. time to fill and empty energy storing elements). In telecommunication the physical delay need sometimes to be modeled and in that case, we may use a ladder of... R and C elements, definitely adding to the confusion!



# Appendix C

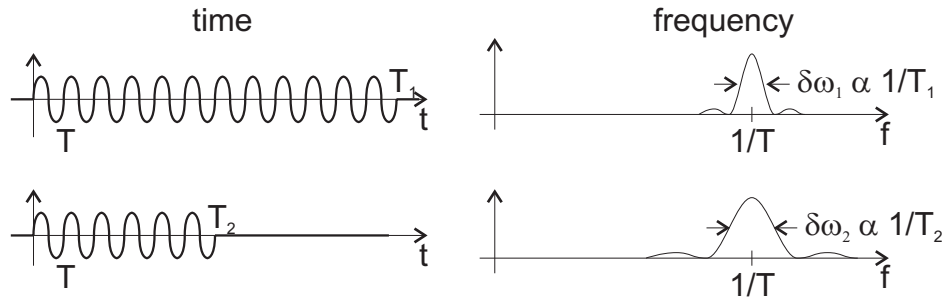
## Resonator and quality factor

The quality factor has originally been introduced to describe oscillators. Actually, the larger the loss you have in an oscillator, the less pure will be its frequency, and thus the poorer its quality.

Simply stated, if a vibrating system (called a resonator, that is an oscillator without circuit to sustain oscillation) has loss, after the oscillations started, the larger the loss, the quicker their amplitude will decrease until they stop. For example, you can think about pinching the string of a guitar: the more loss you have (e.g., if you press the string with a finger on the frets) the shorter the time the vibration will last.

Formally this can be seen by looking at the spectrum of the generated signal. If we had an ideally pure signal it should have only *one* frequency, thus in time domain it could be represented by  $f(t) = A \cos(\omega t)$ . For the resonator it means that it started vibrating at time  $t = -\infty$  and would never end... a quite unphysical signal - the universe after all has only 12 billions years! Thus no oscillator can possibly generate a purely sinusoidal signal, all signals will always have some linewidth  $\delta\omega$ . To understand this we consider that the resonator vibrate sinusoidally between  $t = 0$  and  $t = T_0$  and is stopped before and after these two moments (a slightly more physical signal). The linewidth  $\delta\omega$  can be found from the Fourier's transform of the temporal signal which gives its frequency spectrum:

$$\begin{aligned} f(t) &= \cos \omega t \text{ for } 0 \leq t \leq T_0 \\ &= 0 \text{ otherwise} \\ \Rightarrow \mathcal{F}\{f(t)\} &= \frac{1}{\sqrt{2\pi}} \frac{\sin \omega T_0}{\omega} \Rightarrow \delta\omega \propto \frac{1}{T_0} \end{aligned}$$



Thus as we see here the shorter the time the resonator will sustain the oscillation, the wider its linewidth... and thus, in some way, the lower its quality. We directly observe the relationship between oscillator 'quality' (their narrow linewidth) and the... quality factor, expressed as we know as the ratio of the linewidth over the resonating frequency.

# Appendix D

## Laplace's transform

The table D.1 gives the common properties of the Laplace's transformation.

Properties	Comments
$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t) dt$	Definition
$f(t) = \mathcal{L}^{-1}\{F(s)\}$	Inverse transform
$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}$	Linearity
$\mathcal{L}\{e^{at} f(t)\} = F(s - a)$	s-shifting
$\mathcal{L}\{f(t - a)u(t - a)\} = e^{-as} F(s)$	t-shifting
$\mathcal{L}\{f'\} = s\mathcal{L}\{f\} - f(0)$ $\mathcal{L}\{f''\} = s^2\mathcal{L}\{f\} - sf(0) - f'(0)$ $\mathcal{L}\{f^{(n)}\} = s^n\mathcal{L}\{f\} - s^{n-1}f(0) - \dots - f^{(n-1)}(0)$	t-differentiation
$\mathcal{L}\{tf(t)\} = -F'(s)$	s-differentiation
$\mathcal{L}\{\int_0^t f(\tau) d\tau\} = \frac{1}{s}\mathcal{L}\{f\}$	t-integration
$\mathcal{L}\{\frac{1}{t}f(t)\} = \int_s^\infty F(\sigma) d\sigma$	s-integration
$\mathcal{L}\{f * g\} = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\}$	Convolution $(f * g)(t) = \int_0^t f(\tau)g(t - \tau) d\tau$ $= \int_0^t f(t - \tau)g(\tau) d\tau$

Table D.1: General properties of the Laplace's transformation

The table D.2 gives the unilateral Laplace's transform of some common function (and simultaneously - albeit not really correctly - the inverse transform as well). Note that as the unilateral transform is defined for  $t > 0$  only, the function in the time domain can be considered to be multiplied by  $u(t)$  (that is, the function is

0 when  $t < 0$ ). In this case the bilateral Laplace's transform (where the integral extends from  $-\infty$  to  $+\infty$ ) is the same as the unilateral Laplace's transform. In Engineering, where signal are causal (that is, have an origin in time) the unilateral transform is of course the preferred form, and we often drop the  $u(t)$  function product on the time function, but we still actually imply that the function in time domain is 0 when  $t < 0$ .

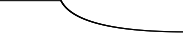

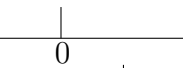
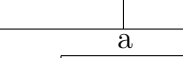

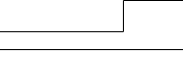
$F(s) = \mathcal{L}\{f(t)\}$		$f(t)$ for $t > 0$
$1/s$		1
$1/s^2$		$t$
$1/s^n$	$(n = 0, 1, 2, \dots)$	$t^{n-1}/(n-1)!$
$1/\sqrt{s}$		$1/\sqrt{\pi t}$
$1/s^{3/2}$		$2\sqrt{t/\pi}$
$1/s^k$	$(k > 0)$	$t^{k-1}/\Gamma(k)$
$1/(s+a)$		Signal decay $e^{-at}$
$a/s(s+a)$		Signal rise $1 - e^{-at}$
$1/(s+a)^2$		$te^{-at}$
$1/(s+a)^n$	$(n = 0, 1, 2, \dots)$	$\frac{t^{n-1}}{(n-1)!}e^{-at}$
$1/(s+a)^k$	$(k > 0)$	$\frac{t^{k-1}}{\Gamma(k)}e^{-at}$
$\omega/(s^2 + \omega^2)$		$\sin \omega t$
$s/(s^2 + \omega^2)$		$\cos \omega t$
$a/(s^2 - a^2)$		$\sinh at$
$s/(s^2 - a^2)$		$\cosh at$
1		Unit impulse (Dirac) $\delta(t)$
$e^{-as}$		Delayed unit impulse $\delta(t-a)$
$1/s$		Unit step at $t = 0$ $u(t)$
$e^{-as}/s$		Unit step at $t = a$ $u(t-a)$

Table D.2: Functional Laplace's transforms

# Appendix E

## Complex numbers

The use of complex numbers in Physics is not ‘forced’, but it simplifies many problems significantly, for example to solve problems in two dimensions, represent waves or periodic signal, ...

A complex number,  $z$  is defined as an ordered pair of real number  $(x, y)$ , where  $x$  is called the real part and  $y$  the imaginary part of the complex number. This complex number can be written in its so-called cartesian form as:

$$z = x + iy$$

where  $i$  is called the imaginary unit and has the property that  $i^2 = -1$ . In Physics  $j$  is often used instead of  $i$ , supposedly to avoid mixing with the symbol used for the current (this is definitely not a good reason, thus, better use  $i$ ). With this simple rule it is possible to use the customary rules of real algebra to compute with the complex numbers.

The possibility of a complex numbers to represent any point in the plane of cartesian coordinate  $(x, y)$  is of extreme importance to solve elegantly problems in two dimensions using algebraic operation, without the need for matrices<sup>1</sup>. Moreover, instead of using the direct orthonormal axes, we may think to use the polar system of coordinate  $(r, \theta)$ ... that gives the polar form of the complex number :

$$z = x + iy = r(\cos \theta + i \sin \theta)$$

where  $r$  is called the modulus (or amplitude, in Physics) of the complex number and  $\theta$  its argument (or phase, in Physics). It is possible to relate nicely the polar form to the complex exponential function by using Euler’s formula  $e^{ix} = \cos x + i \sin x$ , giving another representation of the polar form, very useful in computation.

$$z = x + iy = r(\cos \theta + i \sin \theta) = re^{i\theta}$$

---

<sup>1</sup>There is an interesting mathematical construct that allows to do the same in 3D. These numbers, originally introduced by mathematician W. Hamilton, are called quaternions and have 4 components (and not 3)

The properties of the complex exponential regarding computation ( $e^{z_1+z_2} = e^{z_1} + e^{z_2}$ , ...) and calculus ( $e^z = e^z$ , ...) are the same than the real function.

Going from one coordinate system to the other may seem fairly simple, and it is for the modulus where we always have  $r = |z| = \sqrt{x^2 + y^2}$ . However, great care should be taken when the principal value of the argument is larger than  $\pi/2$  or smaller than  $-\pi/2$ . Let's have a look at Figure E.1 to see what happen,

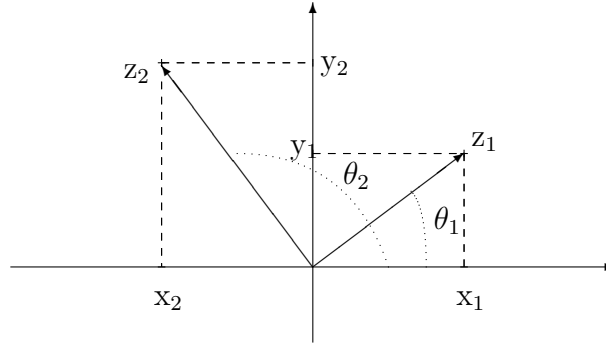


Figure E.1: Complex numbers in different quadrants.

Here we have,

$$\theta_1 = \arctan(y_1/x_1)$$

as expected, but for  $\theta_2$  we have,

$$\theta_2 = \pi/2 + \arctan(|x_2|/|y_2|) \text{ or } \theta_2 = \pi + \arctan(y_2/x_2)$$

showing a very big difference indeed!

If you still want to use a calculator blindly you may use the function for rectangular-polar coordinate transformation. Because you input two data in this case ( $x$  and  $y$ ), and not only one ( $y/x$  or  $x/y$ ) as when you use the arctan function, the calculator is able to know in which quadrant you are, and gives the right answer. However, sketching the complex number in the complex plane won't harm either! Finally, you may always check you have the right angle by verifying that  $x = |z|\cos(\theta)$  and  $y = |z|\sin(\theta)$ . The direct function (sin, cos and tan) don't have the trouble of the inverse ones...

The exponential expression of a complex number is used in Physics to conveniently represent a periodic signal like  $V(t) = V_0 e^{i(\omega t + \phi_0)}$ , where  $\omega$  is the pulsation,  $\phi_0$  the phase at  $t = 0$ , and  $V_0$  the amplitude of the signal. However, it is understood that the only meaningful expression is the real part of the signal  $V(t) = V_0 \cos(\omega t + \phi_0)$  because in standard Physics<sup>2</sup> 'imaginary' numbers have no reality (actually we can also take the imaginary part as the signal, and then forget the real part, the important point is to be consistent in the choice). The use of this exponential notation simplifies tremendously the computation (algebra rules

<sup>2</sup>It may be noted that it is substantially different in quantum mechanics, where the imaginary part of the wave function has a 'real' meaning.

for exponential are simpler than for trigonometric functions) but it can only be used with linear problems. Actually most of the fundamental equations of Physics are linear (Newton's second law, Maxwell's equation, wave equation, diffusion equation, Schrödinger's equation...) and it is often not a serious limitation.

However, for a *non-linear problem* the  $\cos(\omega t + b\phi_0)$  (or  $\sin$ ) function has to be used to represent a periodic signal instead of the  $e^{i(\omega t + b\phi_0)}$ . A simple example will make the reason clear. Imagine a non-linear system that is simply squaring the input signal, i.e.  $y = x^2$ . Note that such system can *not* be described by the theory developed in Chapter 2.4.3.

Now if we use for a periodic input signal the complex signal  $\hat{x}(t) = e^{i\omega t}$ , the complex output becomes  $\hat{y}(t) = e^{i2\omega t}$ , and to obtain the real signal we take the real part of this expression  $y(t) = \Re\{\hat{y}(t)\} = \cos(2\omega t)$ .

If we now use the  $\cos$  directly, we have  $x(t) = \cos(\omega t)$ , thus  $y(t) = \cos^2(\omega t) = \frac{1+\cos(2\omega t)}{2}$ , which is the right answer... but rather different from the first answer!







farther than  $a$ , or alternatively is collimated on the aperture), we may then use the far-field approximation and we talk about Fraunhofer diffraction. In this case the diffraction equation becomes :

$$U_P = \frac{u_A e^{i(\omega t - kR)}}{R} \int \int_A e^{ik(Yy + Zz)/R} dA$$

where  $u_A$  is the amplitude of the illuminating field on the aperture (supposed to be uniform and with a constant phase as it is a plane wave).

Using polar coordinate the aperture area element is expressed as  $dA = \rho d\phi \times d\rho$  and we have  $Yy + Zz = q \sin \Phi \rho \sin \phi + q \cos \Phi \rho \cos \phi = q\rho \cos(\phi - \Phi)$ , thus the field amplitude in P becomes:

$$U_P = \frac{u_A e^{i(\omega t - kR)}}{R} \int_0^a \int_0^{2\pi} e^{ikq\rho \cos(\phi - \Phi)/R} \rho d\rho d\phi$$

The  $d\phi$  integration yields a Bessel function of order 0:

$$\int_0^{2\pi} e^{ikq\rho \cos(\phi - \Phi)/R} d\phi = 2\pi J_0(kq\rho/R)$$

Then the  $\rho$ -integral can be computed as:

$$\int_0^a 2\pi J_0(kq\rho/R) \rho d\rho = 2\pi \frac{R^2}{k^2 q^2} \int_0^{kqa/R} J_0(kq\rho/R) kq\rho/R d(kq\rho/R) \quad (\text{F.1})$$

$$= 2\pi \frac{R^2}{k^2 q^2} kqa/R J_1(kqa/R) \quad (\text{F.2})$$

$$= 2\pi a^2 \frac{R}{kqa} J_1(kqa/R) \quad (\text{F.3})$$

where  $J_1$  is the Bessel function of order 1. Thus the expression of the field amplitude in P is given by:

$$U_P = \frac{u_A e^{i(\omega t - kR)}}{R} 2\pi a^2 \frac{R}{kqa} J_1(kqa/R)$$

Accordingly the irradiance in P is given by:

$$I_P = U_P U_P^* = \frac{2u_A^2}{R^2} \pi^2 a^4 \left[ \frac{J_1(kqa/R)}{kqa/R} \right]$$

Plotting this function with  $kqa/R$  as the argument (Figure F.2) we obtain a central peak with smooth profile surrounded by a series of faint rings. The irradiance becomes zero for  $kqa/R = \pm 3.83$  (first zero of  $J_1(z)$ ), and this point may serve for defining the radius  $Q$  of the illuminated zone on the screen. We get :

$$Q = 3.83 \frac{R}{ak} = 3.83 \frac{R\lambda}{2\pi a} \approx 0.61 \frac{R\lambda}{a}$$

This central peak is called the Airy disk and as suggested earlier, we see that the diffracted spot size (or the diameter of the Airy disk) increases as  $a$  becomes smaller.

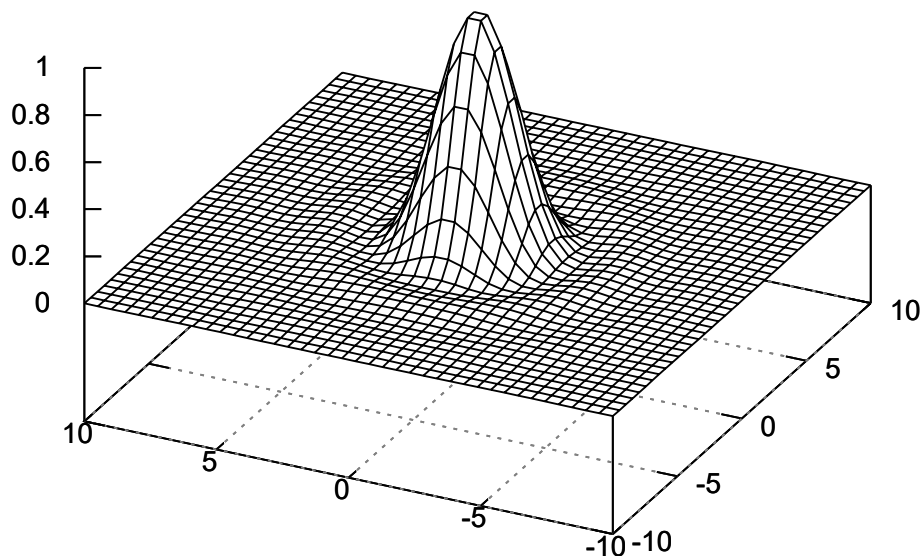


Figure F.2: Intensity profile on a screen located far after a circular aperture uniformly illuminated (scale is normalized to  $kqa/R$ ).

## F.2 Bessel function

Bessel functions are complex functions appearing regularly in physics and not much more complicated than the ‘usual’ cos or sin functions. The Bessel function of order  $n$ ,  $J_n(z)$ , is the solution of the differential equation

$$z^2 \frac{d^2 J_n(z)}{dz^2} + z \frac{dJ_n(z)}{dz} + (z^2 - n^2)J_n(z) = 0$$

that appears often in problem expressed in polar or cylindrical coordinates.

Instead of a differential solution, the Bessel functions are obtained from integral as :

$$J_n(z) = \frac{i^{-n}}{\pi} \int_0^\pi e^{iz \cos \theta} \cos n\theta d\theta$$

or

$$J_n(z) = \frac{1}{2\pi i^n} \int_0^{2\pi} e^{jz \cos \theta} e^{in\theta} d\theta$$

giving for example

$$J_0(z) = \frac{1}{2\pi} \int_0^{2\pi} e^{jz \cos \theta} d\theta$$

and

$$J_1(z) = \frac{1}{\pi i} \int_0^\pi e^{jz \cos \theta} \cos \theta d\theta$$

which are plotted in Figure F.3.

Using numerical solution, we find that  $J_0(z) = 0$  for  $z = \pm 2.40, \pm 5.52, \pm 8.65\dots$  and  $J_1(z) = 0$  for  $z = 0, \pm 3.83, \pm 7.01\dots$

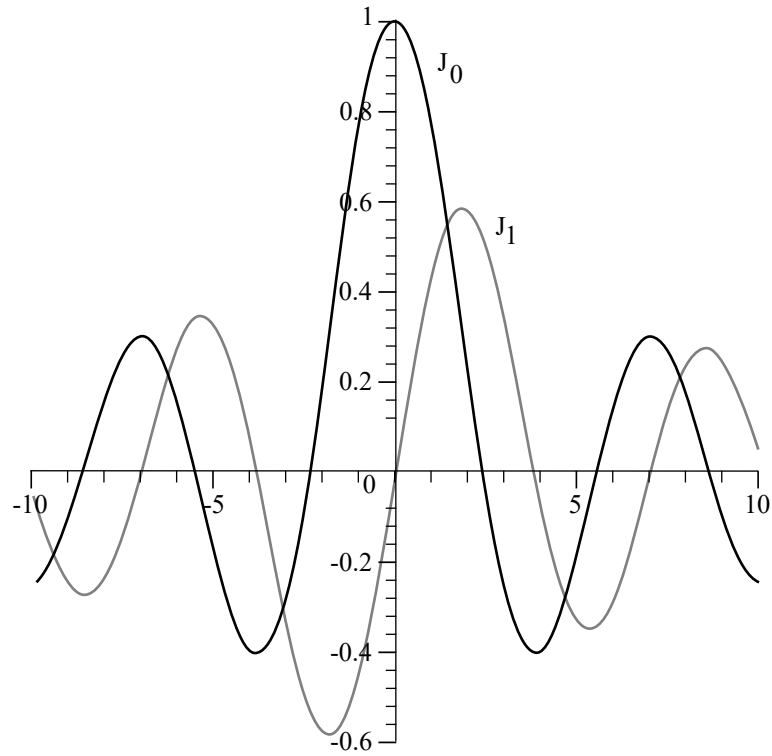


Figure F.3: Bessel function of order 0 and 1.

Moreover, Bessel functions have some interesting properties that can be deduced from their definition that facilitate the resolution of some differential equations. For example we can show that :

$$\left(\frac{1}{x} \frac{d}{dx}\right)^m [x^n J_n(x)] = x^{n-m} J_{n-m}(x)$$

which gives for the  $n = 1$  and  $m = 1$  :

$$\frac{1}{x} \frac{dx J_1(x)}{dx} = J_0(x)$$

or by integration :

$$\int_0^u J_0(v) dv = u J_1(u)$$

# Appendix G

## OCTAVE code

OCTAVE is the free (as in free speech) version of MATLAB<sup>®</sup> and it can be tested directly online at <https://octave-online.net/>. Take care when you copy from PDF file the straight quotes ' get wrongly pasted...

### G.1 Bode diagram

The Bode diagram in Chapter 2.5 have been obtained using the following code, that may be used to plot the Bode diagram of any first or second order transfer function by changing the relevant value.

The first listing is for a first order transfer function.

```
% Bode-plot for first order transfer function
clear all

omega = logspace(-2, 2);      % pulsation range (power of 10)
f = omega/2*pi;              % frequency range

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% First order model %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

G = 1;                        % static gain
tau = 1;                       % time constant

% compute the amplitude
A1 = G./sqrt(1 + tau.^2*omega.^2);

% compute the phase
Phi1 = - atan(tau.*omega);

% plot the Bode diagram
```

```

figure(1)
clf
semilogx(omega, 20*log10(A1))    % amplitude in dB
grid
title(sprintf('Amplitude of a first order system (tau = %i, G = %i)', tau, G))

figure(2)
clf
semilogx(omega, Phi1*180/pi)    % phase in degree
grid
title(sprintf('Phase of a first order system (tau = %i, G = %i)', tau, G))

```

The second listing correspond to a second order transfer function.

```

% Bode-plot for second order transfer function
clear all

omega = logspace(-2, 1, 1000); % pulsation range (power of 10)
f = omega/2*pi;                % frequency range

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Second order system %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
G = 1;                          % Static gain
omega0 = 1;                      % Natural frequency
zeta = 0.1;                      % Damping ratio

% compute the amplitude
A2 = G./sqrt((1 - (omega./omega0).^2).^2 + (2*zeta.*omega./omega0).^2);

% compute the phase
Phi2 = - atan2(2*zeta.*omega./omega0, 1 - (omega./omega0).^2);
% note the use of atan2 instead of atan because the phase exceed
% the pi/2, -pi/2 limits. atan2 takes into account the quadrant in
% which the complex number lies to get the atan in a -pi, pi range.

% plot the Bode diagram
figure(1)
clf
semilogx(omega, 20*log10(A2))    % amplitude in dB
grid
title(sprintf('Amplitude of a 2nd order system (omega0 = %i,...
zeta = %f, G = %i)', omega0, zeta, G))
figure(2)
clf

```

```
semilogx(omega, Phi2*180/pi)      % phase in degree
grid
title(sprintf('Phase of a 2nd order system (omega0 = %i,...
zeta = %i, G = %i)', omega0, zeta, G))
```





# Bibliography

- [1] As recounted by A. Pisano in the foreword of An introduction to microelectromechanical systems engineering, 1<sup>st</sup> edition, N. Maluf, Artech House, Boston (1999)
- [2] J.-C. Eloy, “MEMS market outlook”, Yole Development (2011)
- [3] L.J. Hornbeck and W.E. Nelson, “Bistable Deformable Mirror Device”, OSA Technical Digest Series, Vol. 8, Spatial Light Modulators and Applications, p. 107 (1988)
- [4] C. Smith, “Piezoresistive effect in germanium and silicon”, Physics review, vol. 94, pp. 42-49 (1954)
- [5] J. Price, “Anisotropic etching of silicon with KOH-H<sub>2</sub>O isopropyl alcohol”, ECS semiconductor silicon, pp. 339-353 (1973)
- [6] H. Nathanson, W. Newell, R. Wickstrom, J. Davis, “The resonant gate transistor”, IEEE Transactions on Electron Devices, vol. ED-14, No. 3, pp. 117-133 (1967)
- [7] W. Trimmer, “Microrobot and micromechanical systems”, Sensors and Actuators, vol. 19, no. 3, pp. 267-287 (1989)
- [8] Micromechanics and MEMS - classic and seminal paper to 1990, Ed. W. Trimmer, Section 2 - side drive actuators, IEEE Press, New-York (1997)
- [9] Roark’s Formulas for Stress & Strain, W. C. Young, 6th ed., McGraw-Hill, New-York (1989)
- [10] Fundamentals of Microfabrication, M. Madou, 2nd ed., CRC Press, Boca Raton (2002) : the original (and a bit messy) reference book to the field - what you are looking for is certainly there... but where ?
- [11] MEMS Performance & Reliability, P. McWhorter, S. Miller, W. Miller, T. Rost. Video, IEEE (2001)
- [12] Microsensors - Principles and Applications, J. W. Gardner. Wiley, Chichester, England (1994): Short book that describe all the aspect of microsensor without too many details. A good introduction.

- [13] Semiconductor Sensors, S. M. Sze. Wiley, New-York, USA (1994): In depth description of most microsensors with substantial physics and mathematics coverage, very similar to the ‘Silicon Sensors’ from Middlehoek, with a slightly larger coverage, introducing other semiconductor materials than silicon.
- [14] Microsystem design, S. Senturia, Kluwer, Boston (2001): the still absolute reference book on microsystem modeling, particularly for person with electronics engineering background.
- [15] S. Nagaoka, “Compact latching type single-mode fiber switches fabricated by a fiber micromachining technique and their practical applications”, IEEE J. of Select. Topics in Quant. Electron., vol. 5, no. 1, pp. 36-45 (1999)
- [16] W. Tang, T. Nguyen, R. Howe, “Laterally driven polysilicon resonant microstructures”, in Proceeding IEEE MEMS workshop, pp. 53-59 (1989)
- [17] T. Akiyama, K. Shono, “Controlled stepwise motion in polysilicon microstructures”, J. Microelectromech. Syst., vol.2, no.3, pp.106-110 (1993)
- [18] Fundamentals and Applications of Microfluidics, N-T. Nguyen, S. Wereley, Artech House, Boston (2002): a good book on microfluidics with a lot of details and insight.
- [19] M. Spiga, G.L. Morino, “A symmetric solution for velocity profile in laminar flow through rectangular ducts”, International Communications in Heat and Mass Transfer, vol. 21, no. 4, pp. 469–475 (1994).
- [20] J. Comtois, V. Bright, M. Phipps, “Thermal microactuators for surface-micromachining processes”, in Proceeding SPIE 2642, pp. 10-21 (1995)
- [21] Micromachined transducers sourcebook, G. Kovacs, McGraw-Hill, Boston (1998): one of the first book with a good description of MEMS fabrication technology.
- [22] K.Petersen, “Silicon as a Mechanical Material”, Proceedings of the IEEE, Vol 70, No. 5, pp. 420-457 (1982)
- [23] L. Chen, J. Miao, L. Guo, R. Lin, “Control of stress in highly doped polysilicon multi-layer diaphragm structure”, Surface and Coatings Technology, vol. 141, no. 1, pp. 96-102 (2001)
- [24] Liu H., Chollet F., “Layout Controlled One-Step Dry Etch and Release of MEMS Using Deep RIE on SOI Wafer”, IEEE/ASME Journal of MEMS, vol. 15, no. 3, pp. 541-547 (2006)
- [25] “SU-8: Thick Photo-Resist for MEMS”, Ed. F. Chollet, 19 Sep 2011, <http://memscyclopedia.org/su8.html>

- [26] K. Pister, M. Judy, S. Burgett, R. Fearing, "Microfabricated hinges", *Sensors & Actuators A*, vol. 33, no. 3, pp. 249-256 (1992)
- [27] H. Liu, F. Chollet, "Micro Fork Hinge for MEMS Devices", *Journal of Experimental Mechanics* (special issue: 'Advance in Experimental Mechanics in Asia'), vol. 21, no. 1, pp. 61-70 (2006)
- [28] *Introduction to Microelectromechanical Systems Engineering*, Nadim Maluf and Kirt Williams, Artech House, Boston (1999) : A good introductory book on MEMS
- [29] *MEMS packaging*, T. Hsu, Inspec IEE, London (2004) : The first real book on MEMS packaging with examples - and not the usual IC packaging volume repackaged for MEMS with generalities...
- [30] a reprint of the transcript of the original talk given in 1959 at CalTech appeared in R. Feynman, "There's plenty of room at the bottom", *J. of MEMS*, vol. 1, no. 1, pp. 60-66 (1992) (the paper is available online at <http://www.zyvex.com/nanotech/feynman.html>)



# Index

- actuation
  - electromagnetic, **202**
  - electrostatic, **203**
  - thermal, **212**
- actuator, *see* MEMS actuator
- AFM, *see* atomic force microscopy
- Agilent, 12
- Airy disk, 282
- Alcatel-Adixen, 18, 126
- AlN, 211
- amorphous, **70**, 120
- Analog Devices, 12, 23, 154, 197, 205, 252
- Analogies, 36
- anisotropic etching, 16, **97**
- anisotropy, 70
- annealing, 113
- anodization, 99
- ANSYS, 25
- aperture
  - numerical (NA), 139
  - relative, 140
- AsGa, 211
- aspect ratio, 100
- Assembly, 231
- atomic force microscope, 147
- AZ9260, 130
- beam, 160
- bi-material actuator, 212
- bimetallic actuator, *see* bi-material actuator
- block diagram, 29
- Bode diagram, 43
  - plot, 43
- boron doping, 99
- Bosch, 22, 121, 125
  - process, 104, **127**
- bulk micromachining, 96
- bumping, 249
- calibration, **250**
- capacitive sensing, **196**
- capillarity, 177
- case
  - CERDIP, 236
  - TO, 236
- causality, 271–272
- chemical vapor deposition, 118
  - APCVD, 118
  - LPCVD, 118
  - PECVD, 119
  - UHCVD, 118
- Chemical-Mechanical Polishing, 122
- chip size packaging (CSP), 249
- clean-room, 67
- cleanroom
  - class, 69
- CMP, *see* Chemical-Mechanical Polishing
- coefficient of thermal expansion, 236
- comb-drive actuator, 205
- compensation, 196, **250**, 254
  - actuator, 257
  - explicit, 255
  - implicit, 255
  - monitored, 255
- compliance matrix, 75
- conductance, 243
- conformality, 108
- contact angle, 174

- controller, 31
- Coventor, 24
- crystallographic planes, 71
- CSP, *see* chip size packaging
- CTE, *see* coefficient of thermal expansion
- CVD, *see* chemical vapor deposition
- Czochralsky process, 121
- damping, 47
  - critically damped, 48
  - damping ratio, 47
  - over-damped, 48
  - under-damped, 48
- Deep reactive ion etching, 125
  - ARDE, 128
  - etching lag, 128
  - notching, 128
  - ripple, 127
  - scaloping, 127
- Dektak, *see* stylus profilometer
- Delphi, 12
- depth of field, 138
- design rules, 122, 157
- diaphragm, *see* membrane
- differential sensing, 197
- diffraction, 281
  - Fraunhofer, 282
- diffusion, 111
- diffusion barrier, 110
- DLP, 12, 23, 78, 123, 236, 246
- DRIE, *see* Deep reactive ion etching
- edge-bead, 115
- EDS, *see* energy dispersive X-ray spectroscopy
- electro-osmosis, 10
- electro-osmosis actuator, 207
- energy
  - adhesive, 173
  - cohesive, 173
  - free surface, 172
  - interfacial, 173
- energy dispersive X-ray spectroscopy, 143
- entry length, 183
- environmental scanning electron microscope, 145
- Epitaxy, 120
- Epson, 13
- ESEM, *see* environmental scanning electron microscope
- etch stop, 99
- etching
  - anisotropic, 96
  - dry, 100
  - isotropic, 96
  - wet, 96
- evaporation, 116
  - e-beam, 116
- EVGroup, 18, 100
- eyepiece, *see* ocular
- Fick's laws, 111
- flip-chip, 154
- flow
  - laminar, 182
  - plug, 208
  - Poiseuille, 183
  - turbulent, 182
- flow rate, 185
  - mass, 185
  - volumetric, 185
- foundry process
  - MPW (Bosch), 121, 124, 155–157
  - MUMPS (MEMSCAP), 122, 156, 158
- frequency
  - angular frequency  $\omega$ , 41
  - cut-off, 46
  - natural, 47
  - resonance, 49
- gap-closing actuator, 206
  - pull-in voltage, 207
- getter, 246
- glass, 78
- GLV, *see* Grating Light Valve
- grain, 70
- grating light valve, 12
- heatuator, 201, 212

- hinge, 122, 164
  - torsion, 23
- hydraulic diameter, 187
- hydrophilic, 174
- hydrophobic, 174
- iMEMS, 154
- imprinting, 66
- Institute of Microelectronics, 15
- Intellisense, 24
- ion implantation, 112
- isotropy, 70
- joint, *see* hinge
- Knudsen number, 81, 182
- Laplace's transform, 33
- lattice
  - diamond, 70
  - fcc, 70
- lead frame, 238
- leak rate
  - standard, 243
  - true, 243
- lens
  - relay, 135
- LIGA, 129
- LiNbO<sub>3</sub>, 211
- lithography
  - soft, *see* imprinting
- LOCOS, 110
- look-up table, 253, 257
- low temperature oxide, 118
- LTO, *see* low temperature oxide
- Lucent, 12, 17
- LUT, *see* look-up table
- lyophilic, 174
- lyophobic, 174
- Mach number, 182
- magnetoresistive effect, 200
- magnification, 134
  - angular, 134
  - electronic, 135
  - lateral, 134
- manufacturing accuracy, 21
- market, 13, 14
- mask, *see* photolithography, mask
- material, 73
- mean free path, 81
- membrane, 163
- MEMS, 9
  - actuator, **201**
  - bioMEMS, 12, 16
  - micro-fluidic, 12, **168**
  - optical, 12, 17
  - polymer, 130
  - RF, 12, 14, 17
  - sensor, 12, 14, 16, **193**
- MEMSCAP, 12
- MemsCap, 22, 25, 122
- MemsTech, 15
- micro-world, 10
- microelectronics, 9, 17
  - integration, 153
- microloading, 127
- Microsens, 234
- miniaturization, 10
- Mitutoyo, 141
- model, *see* simulation
  - block representation, 27
  - circuit representation, 27
- MOEMS, 12
- molecular conductance, 243
- Motorola, 12, 236
- MPW, *see* foundry process
- MUMPS, *see* foundry process
- NA, *see* aperture
- natural frequency, 47
- NTT, 203
- numerical aperture, *see* aperture
- ocular, 132
- OMM, *see* Optical MicroMachines
- Optical MicroMachines, 12, 207
- overetch, 122
- Oxford System, 126

- oxidation, **109**
- Péclet number, 182
- parfocality, 135
- Pascal, Blaise (French physicist and mathematician), 80
- pattern generator, 66
- pattern transfer, 64
- patterning, 64
- PentaVacuum, 125
- permeability, 240
- permeation, 240
- Pfeiffer Vacuum, 87
- photolithography, 65
  - mask, 64
- photoresist, 64
  - negative, 65
  - positive, 65
- physical vapor deposition, 116
- piezoelectricity, 200, 208
  - actuator, 209
  - converse effect, 58, 208
  - direct effect, 58, 200, 208
- piezoresistive effect, 16, **193**
- piezoresistor, 195
- Pirani, Marcello (German physicist), 93
- plasma, 101, 116
  - ICP, 126
- Plateau-Rayleigh instability, 176
- POEMS, *see* MEMS polymer
- Poise, 178
- poling, 211
- polycrystalline, **70**, 120
- polymer, 78
- pressure, 79, 80
  - hydrostatic, 81, 90
- process, 63
  - additive, 64, 108
  - back-end, 66, 229
  - front-end, 66
  - modifying, 64, 108, 109, 111
  - subtractive, 64, 96, 100
- process tolerance, 21
- pump, 83
  - cryogenic, 89
  - diffusion, 87
  - Roots, 86
  - screw, 86
  - scroll, 86
  - turbomolecular, 87
- PVD, *see* physical vapor deposition
- pyrolysis, 118
- Q, *see* quality factor
- quality factor, 52
- quartz, 130, 211
- rapid thermal processing, 119
- reactive ion etching, 102
- reflow process, 130
- relative aperture, *see* aperture
- release etch, 123
- reliability, 23, **24**
- resolution, 136
- resolving power, *see* resolution
- resonance frequency, 49
- resonator, 273
- response
  - sinusoidal steady state, 41
  - step response, 41
- Reynold's number, 182
- RIE, *see* reactive ion etching
- RTP, *see* rapid thermal processing
- sacrificial etching, 109
- sacrificial layer, 65, 106, **123**
- Sandia National Laboratory, 77, 122
- scaling laws, 19
- scanning electron microscope, 142
- scanning near-field optical microscope, 142
- scanning transmission electron microscope, 143
- scratch-drive actuator, 207
- SDA, *see* scratch drive actuator
- SEM, *see* scanning electron microscope
- Sensor, 12, 15, 106, 153, 249
- sensor, *see* MEMS sensor
- Sercalo, 12, 77, 205



- shape-memory alloy, 217
- shape-memory effect, 217
- silicon, 73
- Silicon Light Machines, 12
- simulation, 24
  - dynamic, 27
- single crystal, **70**
- SIP, *see* system in the package
- SiTime, 12
- SMA, *see* shape-memory alloy
- SNOM, *see* scanning near-field optical microscope
- SOC, *see* system on chip
- SOI, 22, **77**
- sol-gel, 115
- spectrum, 41, 55
- Spice, 25
- spin-coating, 114
- spin-on-glass, 114
- spring, 160
- sputter, 116
  - DC, 116
  - magnetron, 117
  - RF, 117
- ST Microelectronics, 14
- STEM, *see* scanning transmission electron microscope
- stiction, 124
- stiffness matrix, 74
- Stokes, 178
- Stokes equation, 182, 224
- structural layer, 106
- structure
  - active, 154
  - passive, 154
- STS, 18
- stylus profilometer, 146
- SU8, 130
- superposition theorem, 56
- surface micromachining, 106
- Surface Technology Systems, 126
- surface tension, 171
- suspension, *see* spring
  - folded-beam, 161
- Suss Microtec, 18, 100
- system
  - closed-loop, 30
  - control system, 30
  - first order, 44
  - linear system, 33
  - measurement system, 29
  - non-linear system, 279
  - open-loop, 30
  - second order, 47
  - system order, 33, 271
- system in the package, 153, 231
- system on chip, 231
- Tanner Research, 122
- TEM, *see* transmission electron microscope
- tensor, 74
- Texas Instruments, 12, 15, 23, 24, 78, 123, 207, 246, 252
- TI, *see* Texas Instruments
- time-independent, 33
- tolerance
  - fabrication, 155
  - geometry, 155
  - patterned width, 155
- Torricelli, Evangelista (Italian physicist and mathematician), 90
- transducer, 57
- transducers, 154
- transfer function, 29, 33
- Transistor Outline, *see* case
- transmission electron microscope, 143
- Tronics, 249
- UBM, *see* under-bump-metallization
- under-bump-metallization, 249
- underetch, 96
- Vacuum gauge
  - Bayard-Alpert, 94
  - cold cathode, 95
  - hot cathode, 94
  - ionization, 94
  - membrane, 92

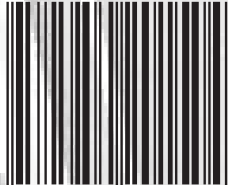
- Penning, 95
- Pirani, 93
- Van der Waals force, 124
- variable
  - effort, 37
  - flow, 37
- velocity
  - gas, 80
  - RMS, 80
- viscosity
  - dynamic, 178
  - kinematic, 178, 188
- wafer bonding, 104
  - thermocompression, 105
- Wafer level packaging (WLP), 249
- wetting, 174
- Wheatstone bridge, 194, 255
- Wheatstone's bridge, 93
- working distance, 141
  
- XactiX, 125
- XeF<sub>2</sub>, *see* xenon difluoride
- xenon difluoride, 125
  
- Young equation, 174
- Young's modulus, 74
- Young-Laplace equation, 176
  
- zeta potential, 208
- ZnO, 211



MEMSCYCLOPEDIA.ORG

ISBN: 978-2-9542015-0-4

ISBN 978-2-9542015-0-4



9 782954 201504