

# The Design of High Performance Mechatronics

This page intentionally left blank

# The Design of High Performance Mechatronics

High-Tech Functionality by  
Multidisciplinary System Integration

Robert Munnig Schmidt  
Georg Schitter  
Jan van Eijk

Delft University Press

© 2011 The authors and IOS Press. All rights reserved.

ISBN 978-1-60750-825-0 (print)

ISBN 978-1-60750-826-7 (online)

doi:10.3233/978-1-60750-826-7-i

Published by IOS Press under the imprint Delft University Press

IOS Press BV  
Nieuwe Hemweg 6b  
1013 BG Amsterdam  
The Netherlands  
tel: +31-20-688 3355  
fax: +31-20-687 0019  
email: [info@iospress.nl](mailto:info@iospress.nl)  
[www.iospress.nl](http://www.iospress.nl)

#### LEGAL NOTICE

The publisher is not responsible for the use which might be made of the following information.

PRINTED IN THE NETHERLANDS

# Contents

<b>Preface</b>	<b>xvii</b>
Motivation . . . . .	xvii
Acknowledgements . . . . .	xix
Summary of the contents . . . . .	xx
<b>1 Mechatronics in the Dutch high-tech industry</b>	<b>1</b>
1.1 Historical background . . . . .	2
1.1.1 The Video Long-play Disk (VLP) . . . . .	3
1.1.1.1 Signal encoding and read-out principle . . . . .	4
1.1.1.2 The Compact Disc and its family members . . . . .	6
1.1.2 The Silicon Repeater . . . . .	9
1.1.2.1 IC manufacturing process . . . . .	10
1.1.2.2 The accurate wafer stage . . . . .	13
1.1.3 The impact of mechatronics on our world . . . . .	16
1.2 Definition and international positioning . . . . .	17
1.2.1 Different views on mechatronics . . . . .	18
1.2.1.1 Cultural differences in mechatronics . . . . .	18
1.2.1.2 Focus on precision-controlled motion . . . . .	21
1.3 Systems engineering and design . . . . .	22
1.3.0.3 Definitions and V-model . . . . .	23
1.3.0.4 The product creation process . . . . .	26
1.3.0.5 Requirement budgeting . . . . .	28
1.3.0.6 Roadmapping . . . . .	29
1.3.1 Design methodology . . . . .	31
1.3.1.1 Concurrent engineering . . . . .	32
1.3.1.2 Modular design and platforms . . . . .	34
<b>2 Electricity and frequency</b>	<b>37</b>
2.1 Electricity and signals . . . . .	38
2.1.1 Electric field . . . . .	38

---

2.1.1.1	Potential difference . . . . .	40
2.1.1.2	Electric field in an electric element . . . . .	42
2.1.2	Electric current and voltage . . . . .	43
2.1.2.1	Voltage source . . . . .	43
2.1.2.2	Electric power . . . . .	45
2.1.2.3	Ohm's law . . . . .	46
2.1.2.4	Practical values and summary . . . . .	47
2.1.3	Variability of electric signals . . . . .	48
2.1.3.1	The concept of frequency . . . . .	49
2.1.3.2	Random signals or noise . . . . .	52
2.1.3.3	Power of alternating signals . . . . .	53
2.1.3.4	Representation in the complex plane . . . . .	54
2.2	Energy propagation and waves . . . . .	57
2.2.1	Mechanical and acoustic waves . . . . .	57
2.2.2	Electromagnetic waves . . . . .	59
2.2.2.1	Transferred energy and amplitude . . . . .	60
2.2.3	Reflection of waves . . . . .	61
2.2.3.1	Standing waves . . . . .	63
2.3	Mathematical analysis of signals and dynamics . . . . .	65
2.3.1	Fourier transform . . . . .	65
2.3.1.1	Triangle waveform . . . . .	67
2.3.1.2	Sawtooth waveform . . . . .	68
2.3.1.3	Square waveform . . . . .	69
2.3.1.4	Fourier analysis of non-periodic signals . . . . .	70
2.3.2	Laplace transform . . . . .	72
2.4	Dynamic system response to a stimulus . . . . .	74
2.4.0.1	Step response . . . . .	74
2.4.0.2	Impulse response . . . . .	76
2.4.0.3	Frequency response . . . . .	78
2.4.1	Graphical representation in the frequency domain . . . . .	79
2.4.1.1	Bode-plot . . . . .	79
2.4.1.2	Nyquist plot . . . . .	83
<b>3</b>	<b>Dynamics of motion systems</b>	<b>85</b>
	Introduction . . . . .	85
3.1	Stiffness . . . . .	87
3.1.1	Importance of stiffness for precision . . . . .	88
3.1.2	Active stiffness . . . . .	92
3.2	Mass-spring systems with damping . . . . .	95
3.2.1	Compliance of dynamic elements . . . . .	95

3.2.2	Combining dynamic elements . . . . .	96
3.2.3	Transfer functions of the compliance . . . . .	101
3.2.3.1	Damped mass-spring system. . . . .	102
3.2.3.2	Critical damping and definition of $\zeta$ . . . . .	106
3.2.3.3	Quality-factor $Q$ . . . . .	110
3.2.3.4	Behaviour around the natural frequency . . . . .	112
3.2.4	Transmissibility of a damped mass-spring system . . . . .	114
3.3	Multi-body dynamics and eigenmodes . . . . .	118
3.3.1	Dynamics of a two body mass-spring system . . . . .	119
3.3.1.1	Analytical description . . . . .	119
3.3.1.2	Multiplicative expression . . . . .	121
3.3.1.3	Effect of different mass ratios . . . . .	122
3.3.2	The additive method with eigenmodes . . . . .	125
3.3.2.1	Multiple eigenmodes and modal analysis . . . . .	129
3.3.2.2	Location of actuators and sensors . . . . .	131
3.3.2.3	Summary . . . . .	135
<b>4</b>	<b>Motion Control</b>	<b>137</b>
	Introduction . . . . .	137
4.1	A walk around the control loop . . . . .	138
4.1.1	Poles and zeros . . . . .	140
4.1.1.1	Controlling unstable mechanical systems . . . . .	140
4.1.1.2	Creating instability by active control . . . . .	141
4.1.1.3	The zeros . . . . .	142
4.1.2	Properties of feedforward control . . . . .	143
4.1.3	Properties of feedback control . . . . .	146
4.2	Feedforward control . . . . .	149
4.2.1	Model based open-loop control . . . . .	149
4.2.2	Input-shaping . . . . .	152
4.2.3	Adaptive feedforward control . . . . .	155
4.3	PID feedback control . . . . .	156
4.3.1	PD-control of a Compact-Disc player . . . . .	157
4.3.1.1	Proportional feedback . . . . .	158
4.3.1.2	Proportional-differential feedback . . . . .	161
4.3.1.3	Limiting the differentiating action . . . . .	163
4.3.2	Sensitivity functions of feedback control . . . . .	167
4.3.3	Stability and robustness in feedback control . . . . .	170
4.3.4	PID-control of a mass-spring system . . . . .	174
4.3.4.1	P-control . . . . .	175
4.3.4.2	D-control . . . . .	176

---

4.3.4.3	I-control . . . . .	177
4.3.5	PID-control of more complex systems . . . . .	182
4.3.5.1	PID-control of a magnetic bearing . . . . .	182
4.3.5.2	Eigenmodes above the desired bandwidth . .	187
4.3.5.3	“Optimal” PID control . . . . .	192
4.3.5.4	Open-loop and closed-loop . . . . .	193
4.4	State-space control representation . . . . .	195
4.4.1	State-space in relation to motion control . . . .	196
4.4.1.1	Damped mass-spring system . . . . .	198
4.4.1.2	PID-control feedback . . . . .	200
4.4.2	State feedback . . . . .	202
4.4.2.1	System Identification . . . . .	204
4.4.2.2	State estimation . . . . .	205
4.4.2.3	Additional remarks on state-space control .	207
4.5	Limitations of linear feedback control . . . . .	209
4.6	Conclusions on motion control . . . . .	214
<b>5</b>	<b>Electromechanic actuators</b>	<b>215</b>
5.1	Electromagnetics . . . . .	217
5.1.0.4	History on magnetism . . . . .	217
5.1.1	Maxwell equations . . . . .	218
5.1.2	Magnetism caused by electric current . . . .	222
5.1.3	Hopkinson’s law . . . . .	225
5.1.3.1	Ferromagnetic materials . . . . .	228
5.1.4	Coil with ferromagnetic yoke . . . . .	229
5.1.4.1	Magnetisation curve . . . . .	230
5.1.5	Permanent magnets . . . . .	231
5.1.5.1	Thermal behaviour and Curie temperature .	235
5.1.6	Creating a magnetic field in an air-gap . . . .	235
5.1.6.1	Optimal use of a permanent magnet . . . .	239
5.1.6.2	Flat magnets to reduce stray flux . . . . .	239
5.1.6.3	Low cost loudspeaker magnet configuration .	241
5.2	Lorentz actuator . . . . .	243
5.2.1	Lorentz force . . . . .	243
5.2.2	Improving the force of a Lorentz actuator .	247
5.2.3	The moving-coil loudspeaker actuator . . . .	248
5.2.4	Position dependency of the Lorentz force . .	248
5.2.4.1	Over-hung and under-hung coil . . . . .	250
5.2.5	Electronic commutation . . . . .	251
5.2.5.1	three-phase electronic control . . . . .	253

5.2.6	Figure of merit of a Lorentz actuator . . . . .	254
5.3	Reluctance actuator . . . . .	257
5.3.1	Reluctance force in Lorentz actuator . . . . .	257
5.3.1.1	Eddy-current ring . . . . .	258
5.3.1.2	Ironless stator . . . . .	259
5.3.2	Analytical derivation of the reluctance force . . . . .	260
5.3.3	Variable reluctance actuator. . . . .	264
5.3.3.1	Electromagnetic relay . . . . .	266
5.3.3.2	Force exerted by a magnetic field . . . . .	267
5.3.4	Hybrid actuator . . . . .	268
5.3.4.1	Double variable reluctance actuator . . . . .	268
5.3.4.2	Combining two sources of magnetic flux . . . . .	271
5.3.4.3	Hybrid force calculation . . . . .	273
5.3.4.4	Magnetic bearings . . . . .	276
5.4	Application of electromagnetic actuators . . . . .	279
5.4.1	Electrical interface properties . . . . .	279
5.4.1.1	Dynamic effects of self-inductance . . . . .	279
5.4.1.2	Limitation of the “jerk” . . . . .	281
5.4.1.3	Damping caused by source impedance . . . . .	282
5.4.2	Comparison of the actuation principles . . . . .	285
5.4.2.1	Standard coil dimension for the comparison . . . . .	286
5.4.2.2	Force of the Lorentz actuator . . . . .	288
5.4.2.3	Force of the reluctance actuator . . . . .	288
5.4.2.4	Force of the hybrid actuator . . . . .	289
5.4.2.5	Dynamic differences . . . . .	289
5.4.2.6	Moving mass . . . . .	290
5.5	Intermezzo: electric transformers . . . . .	291
5.5.1	Ideal transformer . . . . .	292
5.5.2	Real transformer . . . . .	294
5.6	Piezoelectric actuators . . . . .	296
5.6.1	Piezoelectricity . . . . .	296
5.6.1.1	Poling . . . . .	297
5.6.2	Transducer models . . . . .	298
5.6.3	Nonlinearity of piezoelectric transducers . . . . .	301
5.6.3.1	Creep . . . . .	301
5.6.3.2	Hysteresis . . . . .	302
5.6.3.3	Aging . . . . .	303
5.6.4	Mechanical considerations . . . . .	304
5.6.4.1	Piezo-stiffness . . . . .	304
5.6.4.2	Actuator types . . . . .	305

---

5.6.4.3	Actuator integration . . . . .	307
5.6.4.4	Mechanical amplification . . . . .	308
5.6.4.5	Multiple directions by stacking . . . . .	309
5.6.5	Electrical considerations . . . . .	311
5.6.5.1	Charge vs. voltage control . . . . .	311
5.6.5.2	Self-sensing actuation . . . . .	312
<b>6</b>	<b>Analogue electronics in mechatronic systems</b>	<b>315</b>
6.1	Passive electronics . . . . .	317
6.1.1	Network theory and laws . . . . .	317
6.1.1.1	Voltage source . . . . .	317
6.1.1.2	Current source . . . . .	318
6.1.1.3	Theorem of Norton and Thevenin . . . . .	319
6.1.1.4	Kirchhoff's laws . . . . .	320
6.1.1.5	Impedances in series or parallel . . . . .	321
6.1.1.6	Voltage divider . . . . .	322
6.1.1.7	Maximum power of a real voltage source . . . . .	324
6.1.2	Impedances in electronic networks . . . . .	326
6.1.2.1	Resistors . . . . .	326
6.1.2.2	Capacitors . . . . .	328
6.1.2.3	Inductors . . . . .	333
6.1.3	Passive filters . . . . .	335
6.1.3.1	Passive first-order RC-filters . . . . .	335
6.1.3.2	Passive higher-order RC-filters . . . . .	338
6.1.3.3	Passive LCR-filters . . . . .	340
6.1.4	Mechanical-electrical dynamic analogy . . . . .	346
6.2	Active electronics . . . . .	350
6.2.1	Basic discrete semiconductors . . . . .	351
6.2.1.1	Semiconductor diode . . . . .	354
6.2.1.2	Bipolar transistors . . . . .	357
6.2.1.3	MOSFET . . . . .	359
6.2.2	One transistor amplifiers . . . . .	361
6.2.2.1	Emitter follower . . . . .	362
6.2.2.2	Voltage amplifier . . . . .	364
6.2.2.3	Differential amplifier . . . . .	367
6.2.3	Operational amplifier . . . . .	370
6.2.3.1	Basic operational amplifier design . . . . .	370
6.2.3.2	Operational amplifier with feedback . . . . .	372
6.2.4	Linear amplifiers with operational amplifiers . . . . .	373
6.2.4.1	Design rules . . . . .	373

6.2.4.2	Non-inverting amplifier . . . . .	374
6.2.4.3	Inverting amplifier . . . . .	376
6.2.4.4	Adding and subtracting signals . . . . .	377
6.2.4.5	Transimpedance amplifier . . . . .	379
6.2.4.6	Transconductance amplifier . . . . .	381
6.2.5	Active electronic filters . . . . .	383
6.2.5.1	Integrator and first-order low-pass filter . . .	383
6.2.5.2	Differentiator and first-order high-pass filter	385
6.2.6	Analogue PID controller . . . . .	387
6.2.6.1	Transfer function . . . . .	388
6.2.6.2	Control gains . . . . .	390
6.2.6.3	High speed PID control . . . . .	390
6.2.7	Higher-order electronic filters . . . . .	391
6.2.7.1	Second-order low-pass filter . . . . .	392
6.2.7.2	Second-order high-pass filter . . . . .	393
6.2.7.3	Different types of active filters . . . . .	393
6.2.8	Ideal and real properties of operational amplifiers . . .	395
6.2.8.1	Dynamic limitations . . . . .	395
6.2.8.2	Limitations on the inputs . . . . .	401
6.2.8.3	Power supply and output limitations . . . . .	404
6.2.9	Closing remarks on low-power electronics . . . . .	405
6.3	Power amplifiers . . . . .	407
6.3.1	General properties of power amplifiers . . . . .	408
6.3.2	Linear power amplifiers . . . . .	411
6.3.2.1	High output impedance amplifiers . . . . .	413
6.3.2.2	Dynamic loads, four-quadrant operation . . .	417
6.3.3	Switched-mode power amplifiers . . . . .	419
6.3.3.1	First example amplifier . . . . .	419
6.3.3.2	Power MOSFET, a fast high-power switch . .	422
6.3.3.3	Pulse-width modulation . . . . .	424
6.3.3.4	High-power output stage . . . . .	428
6.3.3.5	Preliminary conclusions and other issues . .	432
6.3.3.6	Driving the power MOSFETs . . . . .	432
6.3.3.7	Charge-pumping . . . . .	434
6.3.3.8	Dual-ended configuration . . . . .	435
6.3.3.9	Output filter . . . . .	437
6.3.4	Resonant-mode power amplifiers . . . . .	438
6.3.4.1	Switching sequence of the output stage . . .	440
6.3.4.2	Lossless current sensing . . . . .	444
6.3.5	Three-phase amplifiers . . . . .	444

---

6.3.5.1	Concept of three-phase amplifier . . . . .	445
6.3.5.2	Three-phase switching power stages . . . . .	446
6.3.6	Some last remarks on electronics . . . . .	447
<b>7</b>	<b>Optics in mechatronic systems</b>	<b>449</b>
7.1	Properties of light and light sources . . . . .	451
7.1.1	Light generation by thermal radiation . . . . .	452
7.1.2	Photons by electron energy state variation . . . . .	453
7.1.2.1	Light emitting diodes . . . . .	455
7.1.2.2	Laser as an ideal light source . . . . .	456
7.1.3	Useful power from a light source . . . . .	460
7.1.3.1	Radiant emittance and irradiance . . . . .	461
7.1.3.2	Radiance . . . . .	461
7.1.3.3	Etendue . . . . .	464
7.2	Reflection and refraction . . . . .	465
7.2.1	Reflection and refraction according to the least time . . . . .	465
7.2.1.1	Partial reflection and refraction . . . . .	469
7.2.2	Concept of wavefront . . . . .	470
7.2.2.1	A wavefront is not real . . . . .	471
7.3	Geometric Optics . . . . .	473
7.3.1	Imaging with refractive lens elements . . . . .	473
7.3.1.1	Sign conventions . . . . .	475
7.3.1.2	Real lens elements . . . . .	476
7.3.1.3	Magnification . . . . .	479
7.3.2	Aberrations . . . . .	481
7.3.2.1	Spherical aberration . . . . .	481
7.3.2.2	Astigmatism . . . . .	483
7.3.2.3	Coma . . . . .	485
7.3.2.4	Geometric and chromatic aberrations . . . . .	485
7.3.3	Combining multiple optical elements . . . . .	487
7.3.3.1	Combining two positive lenses . . . . .	488
7.3.4	Aperture stop and pupil . . . . .	491
7.3.5	Telecentricity . . . . .	493
7.3.5.1	Pupil in a telecentric system . . . . .	494
7.3.5.2	Practical applications and constraints . . . . .	495
7.4	Physical Optics . . . . .	497
7.4.1	Polarisation . . . . .	497
7.4.1.1	Birefringence . . . . .	499
7.4.2	Interference . . . . .	501
7.4.2.1	Fabry-Perot interferometer . . . . .	503

7.4.3	Diffraction . . . . .	505
7.4.3.1	Amplitude gratings . . . . .	506
7.4.3.2	Phase gratings . . . . .	508
7.4.3.3	Direction of the incoming light . . . . .	515
7.4.4	Imaging quality based on diffraction . . . . .	515
7.4.4.1	Numerical aperture and f-number . . . . .	519
7.4.4.2	Depth of focus . . . . .	522
7.5	Adaptive optics . . . . .	525
7.5.1	Thermal effects in optical imaging systems . . . . .	525
7.5.2	Correcting the wavefront . . . . .	527
7.5.2.1	Zernike modes . . . . .	528
7.5.2.2	Adaptive optics as correction mechanism . . . . .	532
7.5.3	Principle of operation . . . . .	533
<b>8</b>	<b>Measurement in mechatronic systems</b>	<b>537</b>
8.1	Introduction to measurement systems . . . . .	539
8.1.1	Errors in measurement systems, uncertainty . . . . .	539
8.1.1.1	The ultimate in uncertainty . . . . .	541
8.1.2	Functional model of a measurement system element .	542
8.2	Dynamic error budgeting . . . . .	544
8.2.1	Error statistics in repeated measurements . . . . .	544
8.2.2	The normal distribution . . . . .	545
8.2.3	Combining different error sources . . . . .	547
8.2.4	Power spectral density and cumulative power . . . . .	548
8.2.5	Cumulative amplitude . . . . .	550
8.2.5.1	Variations on dynamic error budgeting . . . . .	551
8.2.6	Sources of noise and disturbances . . . . .	551
8.2.6.1	Mechanical noise . . . . .	552
8.2.6.2	Electronic noise . . . . .	552
8.2.6.3	Using noise data from data-sheets . . . . .	554
8.3	Sensitive signals in measurement systems . . . . .	556
8.3.1	Sensing element . . . . .	557
8.3.2	Converting an impedance into an electric signal . . . . .	558
8.3.2.1	Wheatstone bridge . . . . .	559
8.3.3	Electronic interconnection of sensitive signals . . . . .	565
8.3.3.1	Magnetic disturbances . . . . .	565
8.3.3.2	Capacitive disturbances . . . . .	568
8.3.3.3	Ground loops . . . . .	569
8.4	Signal conditioning . . . . .	571
8.4.1	Instrumentation amplifier . . . . .	571

---

8.4.2	Filtering and modulation . . . . .	574
8.4.2.1	AM with square wave carrier . . . . .	575
8.4.2.2	AM with sinusoidal carrier . . . . .	576
8.5	Signal processing . . . . .	579
8.5.1	Schmitt trigger . . . . .	579
8.5.2	Analogue-to-Digital conversion . . . . .	580
8.5.2.1	Gray code . . . . .	581
8.5.2.2	Sampling of analogue values . . . . .	583
8.5.2.3	Nyquist-Shannon theorem . . . . .	584
8.5.2.4	Filtering to prevent aliasing . . . . .	587
8.5.3	Analogue-to-digital converters . . . . .	587
8.5.3.1	Dual-slope ADC . . . . .	589
8.5.3.2	Successive-approximation ADC . . . . .	590
8.5.3.3	Delta-Sigma ADC . . . . .	593
8.5.4	Connecting the less sensitive elements . . . . .	596
8.5.4.1	Characteristic impedance . . . . .	596
8.5.4.2	Non-galvanic connection . . . . .	598
8.6	Short-range motion sensors . . . . .	599
8.6.1	Optical sensors . . . . .	599
8.6.1.1	Position sensitive detectors . . . . .	600
8.6.1.2	Optical deflectometer . . . . .	603
8.6.2	Capacitive position sensors . . . . .	605
8.6.2.1	Linearising by differential measurement . . . . .	606
8.6.2.2	Accuracy limits and improvements . . . . .	607
8.6.2.3	Sensing to conductive moving plate . . . . .	610
8.6.3	Inductive position sensors . . . . .	611
8.6.3.1	Linear variable differential transformer . . . . .	613
8.6.3.2	Eddy-current sensors . . . . .	615
8.7	Dynamic measurements of mechanical quantities . . . . .	617
8.7.1	Measurement of force and strain . . . . .	617
8.7.1.1	Strain gages . . . . .	618
8.7.1.2	Fibre Bragg grating strain measurement . . . . .	620
8.7.2	Velocity measurement . . . . .	623
8.7.2.1	Geophone . . . . .	624
8.7.3	Accelerometers . . . . .	628
8.7.3.1	Closed-loop feedback accelerometer . . . . .	628
8.7.3.2	Piezoelectric accelerometer . . . . .	630
8.7.3.3	MEMS accelerometer . . . . .	638
8.8	Optical long-range incremental position sensors . . . . .	641
8.8.1	Linear optical encoders . . . . .	642

8.8.1.1	Interpolation . . . . .	646
8.8.1.2	Vernier resolution enhancement . . . . .	648
8.8.1.3	Interferometric optical encoder . . . . .	650
8.8.1.4	Concluding remarks on linear encoders . . . . .	654
8.8.2	Laser interferometer measurement systems . . . . .	656
8.8.2.1	Homodyne distance interferometry . . . . .	657
8.8.2.2	Heterodyne distance interferometry . . . . .	662
8.8.2.3	Measurement uncertainty . . . . .	672
8.8.2.4	Different configurations . . . . .	678
8.8.3	Mechanical aspects . . . . .	684
8.8.3.1	Abbe error . . . . .	685
<b>9</b>	<b>Precision positioning in wafer scanners</b>	<b>687</b>
9.1	Introduction . . . . .	687
9.1.1	The wafer scanner . . . . .	689
9.1.2	Requirements on precision . . . . .	691
9.2	Dynamic architecture . . . . .	695
9.2.1	Balance masses . . . . .	696
9.2.2	Vibration isolation . . . . .	698
9.2.2.1	Eigendynamics of the sensitive parts . . . . .	701
9.3	Zero stiffness stage actuation . . . . .	705
9.3.1	The wafer stage actuation concept . . . . .	706
9.3.1.1	Wafer stepper long-range Lorentz actuator . .	706
9.3.1.2	Multi-axis positioning . . . . .	709
9.3.1.3	Long- and short-stroke actuation . . . . .	710
9.3.2	Full magnetic levitation . . . . .	713
9.3.3	Limits in acceleration of reticle stage . . . . .	714
9.4	Position measurement . . . . .	716
9.4.0.1	The alignment sensor . . . . .	718
9.4.1	Keeping the wafer in focus . . . . .	720
9.4.2	Dual-stage measurement and exposure . . . . .	722
9.4.3	Long-range incremental measurement system . . . . .	723
9.4.3.1	Real-time metrology loop . . . . .	724
9.5	Motion control . . . . .	727
9.5.1	Feedforward and feedback control . . . . .	728
9.5.2	The mass dilemma . . . . .	730
9.6	Main design rules for precision . . . . .	731

---

<b>Appendix</b>	<b>733</b>
Recommended other books . . . . .	733
Nomenclature and abbreviations . . . . .	737
Index . . . . .	746

# Preface

## Motivation

A world without mechatronics is almost as unthinkable as a world without electric light. After its origin around the second world war the name mechatronics has become known for all kind of mechanical systems where mechanics and electronics are combined to achieve a certain function. The complexity of mechatronics ranges from a simple set of electronic controlled relay-switches to highly integrated precision motion systems. This proliferation of mechatronics has been accompanied by many books that each have been written with a different scope in mind depending on the specific technological anchor point of the author(s) within this wide multidisciplinary field of engineering.

The book that you are reading now distinguishes itself from other books in several ways. First of all it is written as a balancing act between both the industrial and the academic background of the authors. The industrial part is based on extensive experience in designing the most sophisticated motion systems presently available, the stages of wafer scanners that are used in the semiconductor industry. The academic part is based on advanced research on ultra precision metrology equipment with fast Scanning-Probe Microscopy and optical measurements with sub-nanometre accuracy. Closely related to the industrial background is the focus on high precision positioning at very high velocity and acceleration levels. With this focus, the book does not include other important applications like robotics and vehicle mechatronics. All presented material is focused on obtaining a maximum of control of all dynamic aspects of a motion system. This is the reason for the term “High Performance” in the title.

A second reason for writing this book next to all others is the observation, when teaching engineering at the university, that most students are rather well trained in applying mathematical rules but too often fail to understand

the full potential of these mathematics in real mechanical designs. The need for the education of real engineers with both theoretical and practical skills, combined with a healthy critical attitude to the outcome of computer simulations, became a guiding motive to finish the tedious job of writing. The capability to swiftly switch between model and reality is one of the most important skills of a real multidisciplinary designer. This capability helps to quickly predict the approximate system behaviour in the concept phase of a design, where intuition and small calculations on the backside of an envelope are often more valuable than computer based detailed calculations by means of sophisticated modelling software. It is certainly true that these software tools are indispensable for further detailing and optimisation in the later phase of a design project but more attention is needed for basic engineering expert-knowledge to cover the concept-design phase where the most important design decisions are taken.

In view of these main motivations to write this book, it was also decided to focus uniquely on the hardware part of mechatronic systems. This means that the important field of embedded software is not presented even though software often serves as the actual implementation platform for modern control systems. The reason for this exclusion is the intended focus of this book on the prime functionality of a mechatronic system, without the interfaces to other systems and human operators. The logical sequence algorithm of the controller, together with the sampling delay, is more important for this prime functionality than the way how this algorithm is described in C-code.

When writing a book on mechatronics, the broad range of contributing disciplines forces a limitation on the theoretical depth to which the theory on each of these disciplines can be treated. Where necessary for the explanation of certain effects the presented material goes somewhat deeper, but most subjects are treated in such a way that an overall understanding is obtained that is based on first principles rather than on specialised in depth knowledge of all details.

Like the work of a mechatronic engineer as system designer in a team of specialists, this book is aimed to be rather a binding factor to the related specialised books than one that makes these redundant.

It is our sincere hope this book serves its purpose.

Also on behalf of the co-authors Georg Schitter and Jan van Eijk,

Robert Munnig Schmidt,

author/editor

July 2011

## Contributions and acknowledgements

Besides much material from our own experience, this book also includes material created by many other people.

Several university staff members and students have contributed to and reviewed the material. Unfortunately it is impossible to mention all without forgetting some names so as example only the three most important students are mentioned.

The first is Ton de Boer, who accepted the impossible task as MSc-student to write the rough material that started this book as lecture notes by following the lectures on Mechatronic System Design. Initially, in spite of professional advice, this writing was done in a well-known WYSIWYG program and only later it was transferred into L<sup>A</sup>T<sub>E</sub>X which proved indeed the only realistic way to create a professional technical textbook.

Leon Jabben and Jonathan Ellis have been working as PhD-students at our laboratory in Delft and parts of their theses are used in the measurement chapter.

Our partners from industry deserve gratitude for their support, financially, in equipment or advice, by permission to use company illustrations or by reviewing the material. The three most important to mention are the Dutch high-tech company ASML and the metrology companies Heidenhain from Germany and Agilent Technologies from the United States.

From ASML especially Hans Butler, Patrick Tinnemans and Jan Mulkens (thanks to the volcano on Iceland!) have helped in reviewing some chapters.

We further thank all other companies and individuals that kindly gave permission to use their illustrations. These all are separately mentioned at the related figures.

It is true to say that this textbook is based on the knowledge of many others as laid down in books, patents and journal articles. For reason of readability we decided not to include references in the text but instead we included a list of the most relevant books that we found to be applicable.

Finally also a word of respect and gratitude should be given to the many contributors of Wikipedia. Even though this huge source of information is not always as consistent and flawless as might be required by the scientific community, Wikipedia has proven to be very useful to quickly find the right physical and mathematical terms or derivations. It also provided information about small trivia like the date of birth or the full name of a famous scientist from the past.

## **Short summary and introduction of the contents**

This book is written in such a way that it is useful both for a high-level student who wants to learn about advanced mechatronics and for engineers in the high-tech industry who want to learn more about adjacent specialisations. To accommodate this dual approach, the first and last chapter determine the environment that makes use of the material of the theoretical chapters in between. It is not a surprise that this first and last chapter are connected by the wafer scanners of ASML as these might well be the most advanced mechatronic systems that are ever designed.

The nine chapters are summarised as follows:

The introduction in Chapter 1 gives the **context of mechatronics in the Dutch high-tech industry** with the historical background, some general observations on the international differences in approach towards mechatronics and the close link with “Systems Engineering”. Subjects include the development of the optical Video Long Play (VLP) disk and the wafer stepper at Philips Research Laboratories. These developments have strongly determined the dominant foothold of high-precision mechatronic system design in the Netherlands and are exemplary for the specific photon-physics oriented approach in this country, so quite different from the machining oriented approach in most other countries. The overview on systems engineering and design introduces some functional design and development methodologies that have proved to be crucial for the success of the high-tech industry. These methods are based on industrial practice where complex multidisciplinary designs have to be realised. Systems Engineering is a field closely related to mechatronics and the corresponding principles are used in structuring the design of a mechatronic system.

Chapter 2 is the first of a series of chapters on the basic theory that is applied in controlled motion systems. It consists of a short overview of the principles of **electricity, frequencies, waves and signal responses**. The chapter starts with basic electricity, the linking element in a mechatronic system. Followed by signal theory this chapter explains the reason why the properties of mechatronics are so often described in the frequency domain next to the more mechanical oriented time-related step and impulse responses. The chapter also introduces different graphical representations of these responses as this material is used throughout most chapters in this book.

The hard-core of a mechatronic system is still the mechanics that represent the real, dynamic, hardware world that has to be mastered when positioning objects in a controlled way. In most cases, the dynamic properties of

the mechanical construction determine the control performance. Expert knowledge of this field is a prerequisite for a mechatronic designer. For that reason Chapter 3 deals with these **dynamics of motion systems** and mainly concentrates on the uncontrolled properties of standard mechanical elements consisting of a multitude of springs, masses and dampers. As a first step towards active motion control this theory enables to determine dynamic causes for observed instability issues in controlled motion systems. Immediately related to the mechanical dynamics is the important field of **active motion control** in Chapter 4. This chapter concentrates on a thorough understanding of the working principle and tuning of the still widely used PID controllers. Also a short introduction is given in more modern model-based control approaches that are expected to play an increasing role in mechatronic systems. A strong emphasis is put on the insight that control both adds virtual elements from the mechanical domain like springs and dampers and new elements like integration.

**Electromechanic actuators and analogue electronics** are two closely related hardware components of a mechatronic system. Their interaction is increasingly underestimated by system designers, because of two reasons. Firstly the field is controlled by experts in physics and electronics. These specialists have a fundamentally different more abstract frame of view than the mostly concrete-mechanical visually oriented system designers. The second reason for underestimating these related fields is caused by the overwhelming amount of electronics and electro-motors that are around us, giving rise to the idea that their principle is simple and mastered by many. This idea is a dangerous delusion as the difficulty in electronics is related to its dynamic analogue behaviour and unfortunately the number of people that master that part is rather decreasing than increasing. It is the analogue side of electronics that deals with measurement and actuation that needs most of the attention of the mechatronic designer.

With this purpose in mind Chapter 5 first presents **linear electromechanic actuators**. This chapter mainly focuses on electromagnetic actuators but also piezoelectric actuators are presented as these are increasingly applied in precision mechatronic systems. This chapter will help in the selection process of actuation systems and creates a knowledge base for further study on the subject. Also the relation with power-amplifier constraints, that are presented in the following chapter, is made clear.

Chapter 6 deals with **analogue electronics** for measurement and power and starts at a very basic level with passive components because most mechanical engineering students have hardly any knowledge about electronics.

The introduction of the active components leads to their application in the basic design of the operational amplifier, the most universal and widely used analogue electronic building block. The last section in this large chapter gives an overview of the basic design of **Power Amplifiers** that act as the interface between the controller and the actuators.

Optics has become a main driver of mechatronic advancement in the past decades and for that reason Chapter 7 gives an introduction to **optics** from the perspective of a mechatronic designer. Optics are important in two ways. Firstly it is an application area where mechatronics are used to control and correct optical properties of imaging systems and other instrumentation. Secondly optics are used to determine distances in a plurality of sensors, that enables us to create measurement systems with extreme precision. Starting with basic physics on optics with sources and the duality of light, an overview of geometrical and physical optics is presented including limiting factors for the performance of imaging systems. The chapter concludes with an introduction on adaptive optics.

Chapter 8 presents the basic principles of **sensors for force and dynamic position measurements** based on several physical principles including strain-, inductive-, capacitive- and optical sensors. The theory in this chapter will enable the first selection of suitable sensors when designing a mechatronic system. Laser interferometry and encoders will also be presented as these are most frequently applied in high precision mechatronic systems. Even though metrology in general will be shortly touched, the chapter concentrates on measurement for control. For this reason also the principle of Dynamic Error Budgeting is included, a statistical method to determine the total error in a dynamic precision system from contributions of different error sources.

As closure of the book Chapter 9 presents the mechatronic design for **precision positioning in waferscanners** where all theory is applied to its most extreme level. This chapter includes the basic design of positioning stages, the need for and active control of vibration isolation, and the motion control approach to achieve a position accuracy of less than a nanometre at speeds of more than 1 m/s and accelerations of more than 30 m/s<sup>2</sup>.

# **Chapter 1**

## **Mechatronics in the Dutch high-tech industry**

This introductory chapter places the subject of this book in the context of the rapid development of the high-tech industry in the Netherlands. This industry has become a main driver of the economic growth in the industrial region around Eindhoven with the multinational company “Royal Philips Electronics” as the most prominent original source of technical innovation. Based on their wide market scope with both consumer and industrial products Philips gave birth to several high-tech spin-off companies among which ASML would become the global market leader in exposure equipment for semiconductor manufacturing. The first part of this chapter will pay tribute to the achievements within the research and development departments of “Philips Gloeilampen fabrieken”, as the company was named at that time, because these have determined the success of high performance mechatronics in both high-tech and consumer equipment. After this short historical overview, the second section will position the field of mechatronics on the international playground where precision machining and manufacturing appear to be the main application fields. The last section deals with systems engineering as a design and development framework for the highly complex mechatronic systems that are applied in the high-tech industry. It will also give a general overview on the related development processes.

## 1.1 Historical background

High performance mechatronics originated in the Netherlands around 1970 – 1980 at Philips Research Laboratories near Eindhoven. Although already world famous at the time, this laboratory was humbly called the “Nat-Lab”, a name that was based on the original Dutch name “Natuurkundig Laboratorium”. The research activities were characterised by multidisciplinary teamwork on subjects with an estimated large application potential. Even though frequently this potential was far from proven and might take a very long time to reach reality, research people with vision and ambition got the opportunity to work on their dream.

The multidisciplinary character of the research subjects was especially evident in the Optical Research Group where in those early days a unique combination of people worked together in close harmony on the optical *Video Long Play Disk* (VLP) and the *Silicon Repeater*. These people were not consciously aware of the fact that they were pioneering in a new field of expertise. A field that later would be known under the name “mechatronics”.



**Figure 1.1:** The Philips “Nat Lab” research laboratories in Eindhoven where the real breakthrough in mechatronics was initiated.

(source: Philips Technisch Tijdschrift Vol.43 nr 2/3/4)

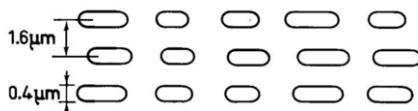


**Figure 1.2:** The VLP Optical disk player had a contact-less optical read-out principle with light from a Helium-Neon laser that was focused on the track of the disk.

(source: Philips Technisch Tijdschrift Vol.43 nr 2/3/4)

### 1.1.1 The Video Long-play Disk (VLP)

In those days, only the gramophone record player and cassette recorder were known for music playback and home registration, while for video only the video cassette recorder (VCR) existed. Even though pre-recorded videotapes would eventually become very popular, the general opinion was at that time that tapes would always be too expensive for the consumer market, due to the large number of parts and expensive material in one tape-cassette. There were already several people in the world with the vision that a gramophone record for video would fundamentally solve that problem, because moulding a gramophone record is not costly at all. Unfortunately, the registration of video images in analogue technology would have required a bandwidth of more than 4 MHz for normal Television signals and that is more than a hundred times higher than needed for the reproduction of sound with a maximum frequency of 20 kHz in two-channel stereo. Based on the fact that even this 20 kHz was quite difficult to achieve with a contact stylus with diamond tip, including the unavoidable contamination of the grooves, it was not considered possible to register video in a comparable way! The breakthrough in thinking, that was required for the future success, was based in the understanding that the contact method would need to be replaced by a contact-less read-out principle with light. With the then newly developed laser as a light source and the application of precision optics, a far higher density of information could be registered than would ever be possible



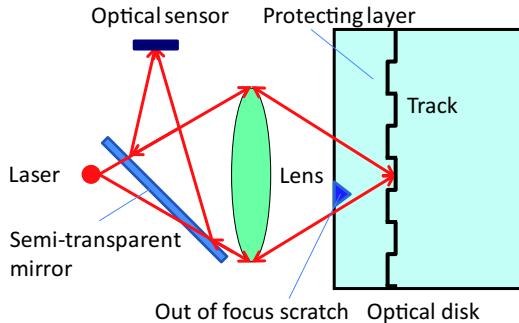
**Figure 1.3:** The signal on a VLP disk is coded in pits on a spiral wound track with a variable length and frequency at a fixed width and mutual distance. The frequency of the pits determined the video signal and the length of the pits determined the audio signal.

by mechanical means. The benefit of contact-less read-out of a rotating disk, as shown in Figure 1.2, becomes even more clear with a small calculation on the registration of high-end audio with a 22 kHz signal on a gramophone record that rotates with 33 rounds per minute. If the signal is located on the inner groove of the disk at a diameter of 130 mm, one period of this signal corresponds with a track-length on the disk of approximately 10  $\mu\text{m}$ . The tip radius of the diamond stylus can not be made much smaller than this same order of magnitude, because otherwise the lifetime would be too short. The tip radius effectively limits the detection of a shorter wavelength. Even with the most refined methods, as are presently applied in the recently revived gramophone players, the maximum frequency that can be registered on a gramophone record is limited to about 40 kHz. With suitable optics and a Helium Neon Laser source with a wavelength of 633 nanometre it was found to be possible to detect details on a rotating disk with just less than a micrometre. Compared with the 10  $\mu\text{m}$  of the mechanical record player, this value is less than one hundredth of the surface area per detail, which is sufficient for the registration of an hour of video information on a disk with the size of a normal gramophone record.

### 1.1.1.1 Signal encoding and read-out principle

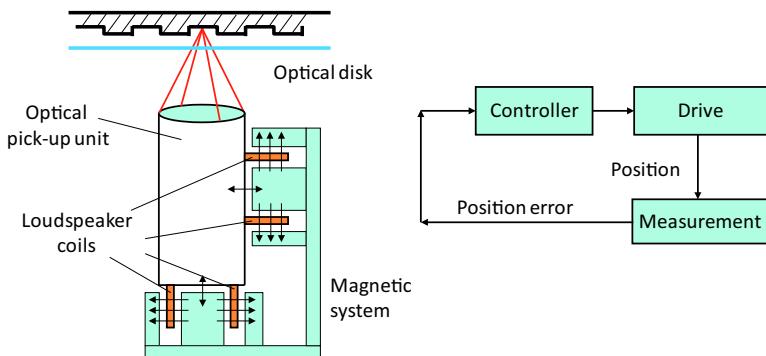
For reason of efficiency, it was decided to directly write the analogue video signal on the disk as a series of pits, tiny stripes with a variable frequency and length. While the frequency contained the video information, the sound was registered in the length of the pits as shown in Figure 1.3.

The pits were written in a spiral-wound track with a width of 0,4  $\mu\text{m}$  at a mutual distance of 1,6  $\mu\text{m}$  on a transparent disk with the same size as an ordinary gramophone record, a diameter of 0.3 m. The rotating speed would be a factor ten higher however, with up to 30 revolutions per second, in order to be able to handle the high frequency. Also different to



**Figure 1.4:** The basic optical read-out principle of an optical disk by focusing a laser beam on the track that is embedded in the optical disk and measuring the reflected intensity with an optical sensor. A scratch on the surface of the protecting layer is not detected because it is not in focus.

the gramophone record was the location of the track, that no longer was embedded in a groove, but was hidden inside the disk and detected by means of a small spot of light through the transparent covering layer, as shown in Figure 1.4. This really appeared to be a revolutionary way of thinking. Some of the competitors, like RCA, still tried to read the information written in a groove on the surface with the help of a stylus. In this case the stylus was only used to follow the track, while the high frequency signal was detected with a local capacitive sensor on the tip of the stylus. With this splitting of functions, following and detecting, they did succeed in getting the mechanical system operational and it has even been some time on the market with well protected disks in cassettes to prevent contamination. This product was in itself a significant achievement for that time but when compared with the contact-less optical read-out system of Philips it could not survive. Especially the lack of sensitivity for surface scratches appeared to be one of the big advantages of optical detection. The scratches are not detected because they are not in the focal point of the detecting laser beam. A problem with this principle was however the fact that no mechanical means were available to follow the track and this created the need to look for ways to actively control the position of the *optical pick-up unit* that was designed to read the information on the disk. It was the solution to this problem that determined the crucial breakthrough of precision mechatronics in the Netherlands. From the signal of a special segmented photo-diode with some additional optics, the relative position of the optical pick-up unit to the track could be measured and with tiny linear moving-coil motors, similar to the ones that are used in loudspeakers, it became possible to keep the



**Figure 1.5:** The active position control of the optical pickup-unit by measuring the position from information in the reflected light and correcting position errors by means of moving coil actuators.

pick-up unit right on track by means of a real fast position control system, as shown in Figure 1.5. This so called *servo-system* was capable to correct the different disturbances and imperfections in the track. Today this correction is achieved even in the cheapest CD-players with special digital processors in one IC that perform all the control tasks, but at that time this control system could only be realised with operational amplifiers, RC-networks and adjustable resistors.

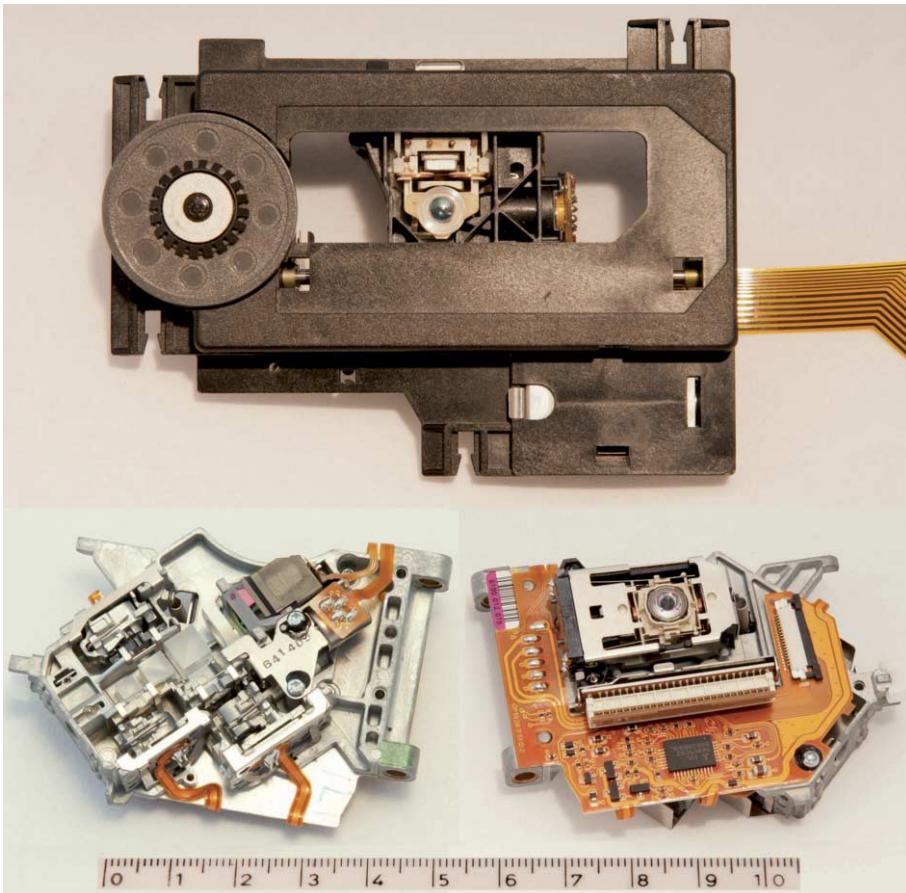
### 1.1.1.2 The Compact Disc and its family members

The VLP was the first product that was brought on the market with optical recording but in spite of the cooperation with MCA for the movie software, the introduction under the name of “Laservision” was eventually not successful. This was mainly caused by the lower than envisaged cost of the video cassette, while the VCR-player could also record video signals directly from television. The missing, at that time not sufficiently developed, possibility to record optical disks on a consumer product, appeared to be the most important drawback of the Laservision disk. This problem would be solved some time later with recordable CDs but that came only after the second breakthrough that reached almost every human being in the world, the invention of the audio *Compact Disc player* of which one of the first commercial products is shown in Figure 1.6. With many successful and less successful successors like CD-Interactive, CDROM, CDRAM, DVD Superaudio CD and Blu-Ray disk, this development really has made a major difference to the entertainment industry. With the CD three major developments were

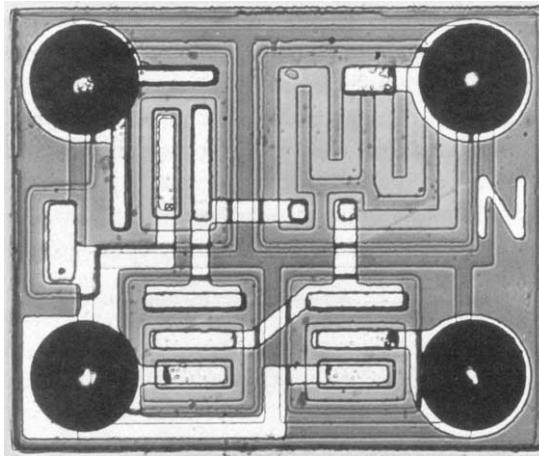


**Figure 1.6:** The CD 303 was the first real high-end CD player of Philips. The complete optical read-out mechanism with the rotating motor for the CD and the rotary arm that carried the optical pick-up unit, was built into a separate drawer that would enable to load the disk from the front. This innovation would not have been easily possible with a normal gramophone record player.

united, the digitisation of analogue signals, the optical disk principle of the VLP and the invention of the semiconductor laser, that enabled a far smaller design of the optical system. As an example of the miniaturisation, that accompanied these developments, the complete mechanism of a CD-player and the optical pick-up unit from a Blu-ray player are shown in Figure 1.7. The design principles that are applied in these systems with the long-stroke, short-stroke splitting of precision and range will return in the last chapter of this book that describe the more recent developments in wafer scanners that started as described in the following section.



**Figure 1.7:** The optical pick-up unit in the CD-mechanism on top of this picture is carried by a simple and inexpensive linear moving stage that is driven with a rotating motor and a screw-nut/gearwheel transmission. This long-stroke linear drive roughly positions the pick-up unit to within the capture range of the precision track-following servo-system with the moving-coil motors of the pick-up unit. This basic long-stroke, short-stroke positioning concept is refined in later developments like the Blu-ray disk of which the optical pick-up unit is shown below the CD-mechanism. This unit is designed to be compatible with all previous CD and DVD related formats. For that reason three different laser sources are integrated in one unit, working at three different wavelengths while all light is optically combined to one spot by means of several semi-reflecting cubes and lenses.



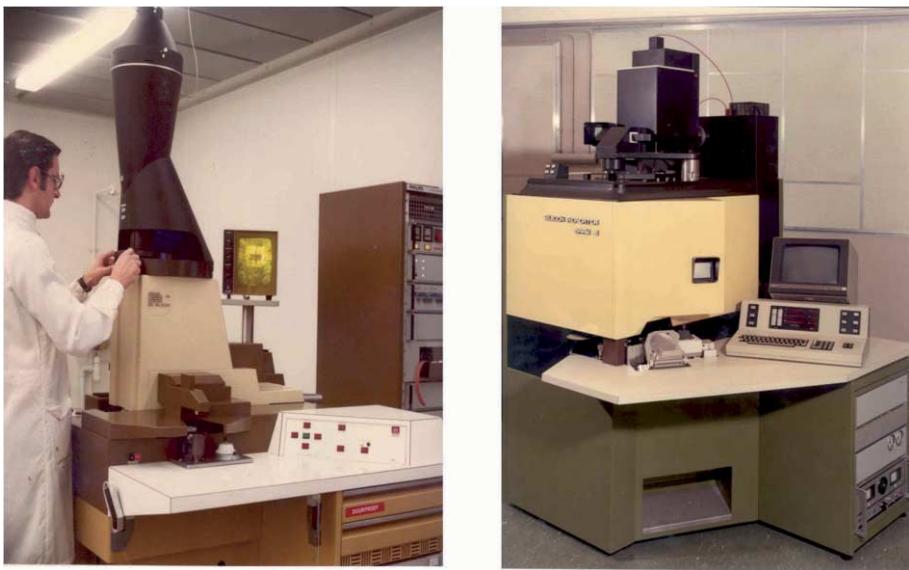
**Figure 1.8:** The OM 200 was the first integrated circuit made by Philips in 1965. The size was  $0.75 \times 0.75 \text{ mm}^2$  and it integrated three transistors and two resistors.

(source: Philips Technisch Tijdschrift Vol.43 nr 2/3/4)

## 1.1.2 The Silicon Repeater

Around the same time as optical recording was invented, Philips produced its own semiconductors. To support this activity, within the Natlab several research groups were active in the global battle to decrease the size of the details in semiconductor based *Integrated Circuits* (ICs). The first commercial integrated circuit of Philips is shown in Figure 1.8 with only a very small number of components, but gradually the number of components increased dramatically with sometimes millions of transistors, resistors and capacitors that have to be realised on a single silicon substrate. These electronic elements are all connected by a multitude of different layers with wiring on top of the active elements, the lighter regions in Figure 1.8.

The technical support and design group within the Natlab supported this research with in-house designed and manufactured precision equipment, like the “Opticograph” that could write a mask for an IC by means of a scanning focused light beam. Among these machines, the *Silicon Repeater*, of which two generations are shown in Figure 1.9, was designed for exposing the pattern of an integrated circuit. The first generation (Mark 1) was designed together with the engineering department of Philips Semiconductor in Nijmegen, while Silicon Repeater Mark 2 was purely made inside the Natlab.



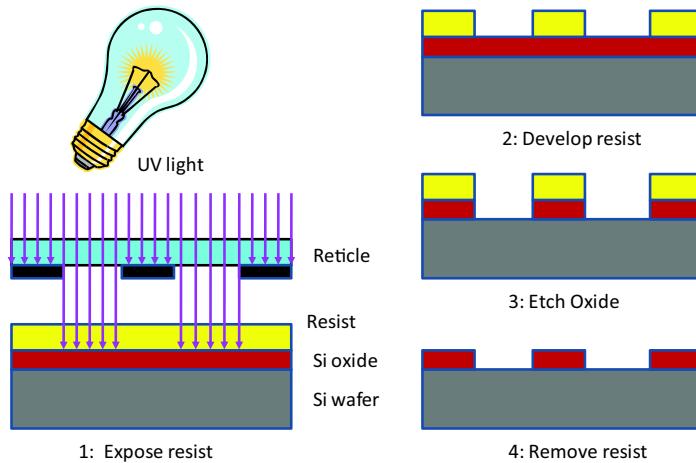
**Figure 1.9:** The Silicon Repeaters Mark 1 and Mark 2 were the first wafer steppers of Philips and achieved an imaging resolution in the micrometre range. They were designed as laboratory tools to enable advanced research on IC production technology.

(source: Philips Technisch Tijdschrift Vol.37 nr 11/12)

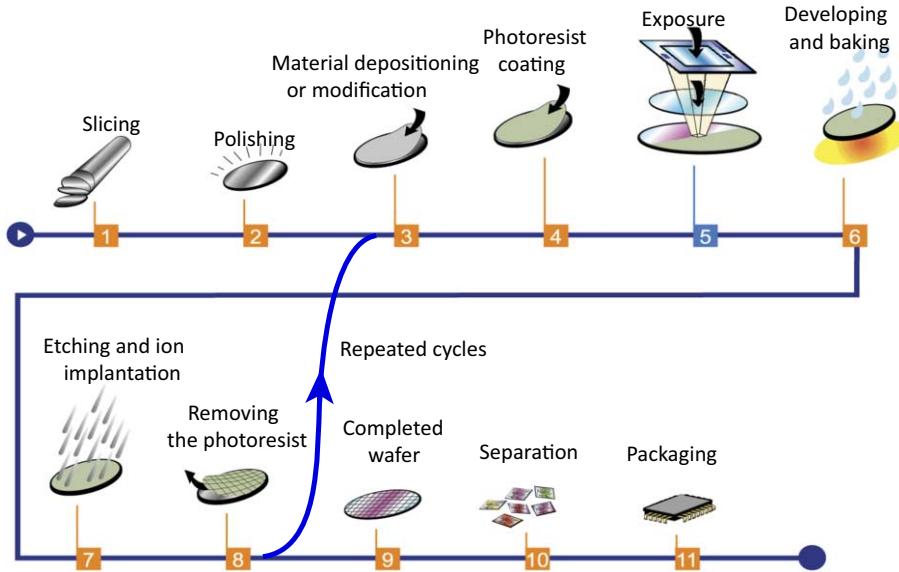
### 1.1.2.1 IC manufacturing process

Both for the realisation of the active electronic components in the IC and the related wiring, it is necessary to work the silicon by means of etching, oxidising or changing the properties locally by different chemical elements. The related process is called lithography which name is based on the ancient Greek words “λιθος” = stone and “γραφειν” = writing.

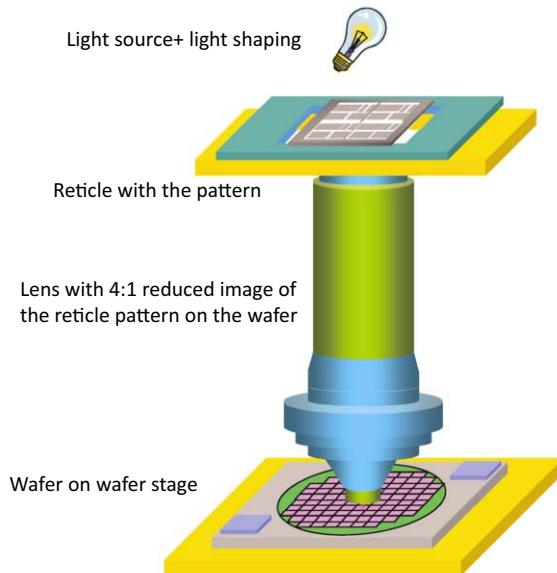
The first step in this process is to “print” the desired pattern by means of an optical exposure of a light sensitive resist layer. This exposure is done by an optical system that images the mask (also called a *reticle*) on the resist. Subsequently, after the development of the resist, the silicon can be chemically treated on the spots where the resist was illuminated, as shown in Figure 1.10. It is essential to note that the same step has to be repeated up to around thirty times, where each different layers has to be defined in one exposure cycle. All exposed layers have to be positioned with a high accuracy relative to the previous layers. In Figure 1.11 a schematic overview of this IC manufacturing process is shown.



**Figure 1.10:** The lithographic process is a local chemical treatment of the silicon substrate. The treated areas are determined by a resist layer that contains a pattern that is previously imaged from the pattern on a mask by an optical exposure system.



**Figure 1.11:** IC production flow starting with a grown mono-crystalline Silicon ingot cut into thin slices (wafers) that undergo a multitude of chemical treatments. The details of the integrated circuit are determined by the optical exposure system at step 5.



**Figure 1.12:** Schematic drawing of the main components of a wafer stepper. The mask is imaged on the wafer with a demagnifying lens. Due to the high opening angle that is required for the high resolution only small areas can be exposed at one exposure at the same time. This requires the wafer to be exposed in steps by means of a wafer stage.

The smallest details of an IC determine both its energy consumption and functionality. Together with the continuous increase in productivity of the manufacturing process for cost reduction, these factors are the most important economic drivers in the semiconductor industry. Both factors are mainly determined by the exposure system. Though all other process steps also need to be capable of realising a high level of refinement, the exposure step is the only one that deals with the detailing of the circuits. For this reason the exposure step has been of overriding importance for the developments in the semiconductor industry. Initially the exposure was done by direct illumination through a mask that was closely positioned above the wafer, like shown in Figure 1.10. This *shadow mask* exposure principle is however limited in resolution by the laws of diffraction and suffers from vulnerability of the previous layers for damage by touching the surface of the mask. To solve that issue, the wafer stepper and the wafer scanner were invented.

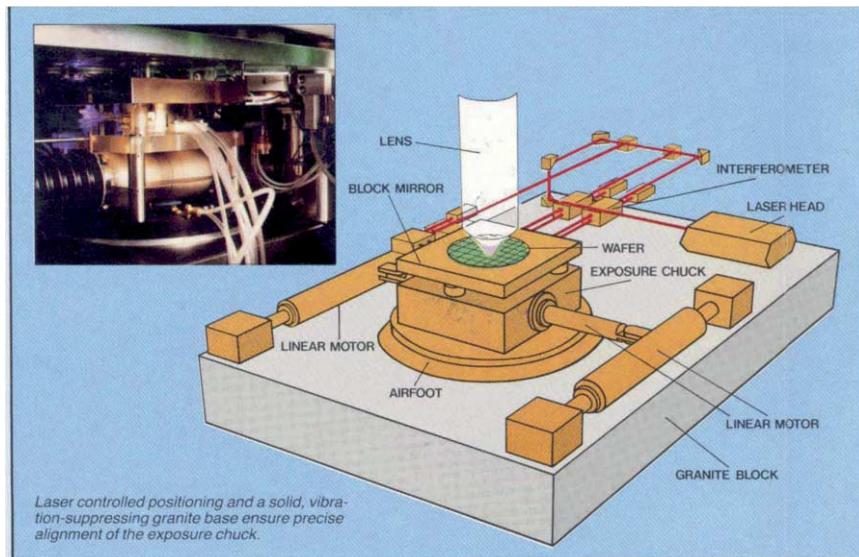
The main components of a wafer stepper are shown in Figure 1.12. The optical projection system is located in the centre. Like a slide projector,

that projects light through a transparency and a lens to the wall, a wafer stepper projects light through a reticle and lens to a silicon wafer. The only difference is the demagnification of the image on the silicon wafer which is related to the required high resolution. To realise a high resolution, the opening angle of the lens needs to be very high as will be explained in Chapter 7 on optics. For this reason the image has to be very close to the lens and light should come from all directions. Exposing the entire wafer in one step would require a lens with a very large diameter which would be too expensive to manufacture. It was concluded that only a small area could be exposed in one exposure at the same time. This means that the wafer has to be exposed in steps while it should be positioned very accurately between the different exposures by means of a mechanism, the *wafer stage*. This stepping motion has to be done extremely fast to not lose time and keep the productivity on an acceptable level. For that reason the speed of this highly accurate movement became one of the key aspects that drove the need for perfection of mechatronics in these machines.

### 1.1.2.2 The accurate wafer stage

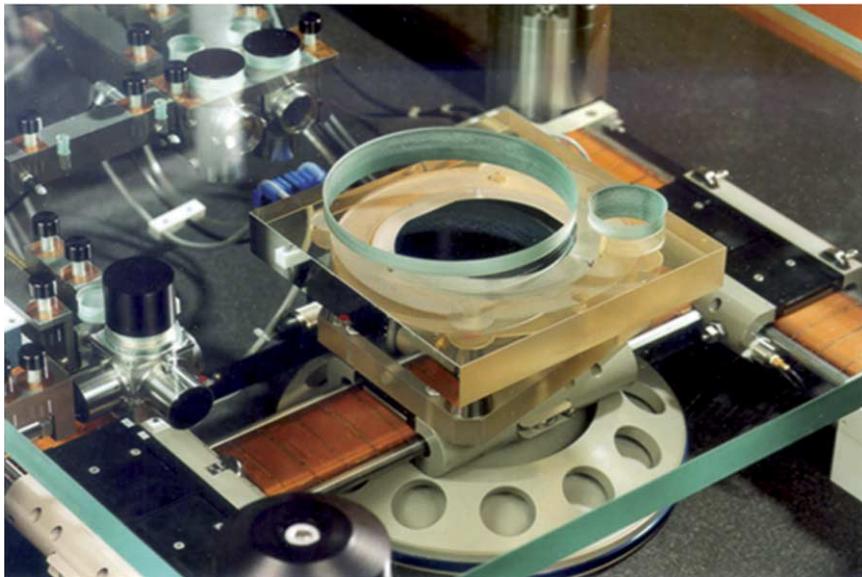
The most important and difficult part from a mechatronic perspective was and still is the wafer stage. The accuracy of the positioning of the wafer in the Silicon Repeater was in the same order of magnitude as the accuracy of the previously described Video Long-play Disk with an error of significantly less than a micrometre over a range of several tens of centimetres. Instead of following a fast moving track, the wafer in a waferstepper had to be positioned in steps followed by a perfect stand-still during exposure. This standing-still had to be realised with a position repeatability of better than  $0.1 \mu\text{m}$  on a wafer of at that time only 5 inch ( $\approx 12 \text{ cm}$ ), which is a better accuracy than one over a million.

The first wafer stages were driven by *hydraulic linear motors* as schematically shown in Figure 1.13. Precision hydraulics was a known field of expertise within Philips, originating from the design of high precision machining tools that were needed in the laboratory to manufacture for instance optical parts. Hydraulic cylinders and *servo-valves* were a proven technology for high precision positioning under several strict conditions. Friction had to be avoided in the hydrostatic bearings and the temperature of the hydraulic oil had to be actively controlled to around  $100 \text{ mK}$  in order not to influence the measurement of the position by laser interferometers. With this wafer stage a repeatability of  $0.1 \mu\text{m}$  was realised by using special valves that control the differential pressure over the hydraulic motors without any back-



**Figure 1.13:** The first wafer stages had hydraulic linear motors to achieve the required stiffness, speed and accuracy. (source: commercial leaflet Philips)

lash by internal friction. In spite of the good performance, hydraulic drives with the necessary high pressure pumps and oil sweating sealings were not the most ideal mechanical parts to be applied in a clean room of an IC factory. One incident with a bursting high pressure tube on a prototype in the laboratory made everyone aware of the problems related to clean everything again. For that reason an electric alternative was perceived to be a far better solution and in parallel, research was done on an alternative electric stage, driven with rotating servo-motors and a friction-wheel transmission. The prototype that was built with this principle showed problems with dynamics due to the transmissions and low stiffness connections and eventually a direct-drive method provided the real replacement of the hydraulic drive. The design of this direct-drive wafer stage was based on the research on the Video Long-play Disk with its magnetically suspended optical pick-up unit. It was recognised that those same principles could also be applied to larger positioning systems, notwithstanding the intuitive drawback of the large mass. At that time linear motion was mostly achieved by screw spindles and rotating motors like the example with friction wheels. In principle the position repeatability of such indirect drives is determined by the manufacturing precision of the parts and generally these are not accurate enough for micrometre accuracy. By applying a position control system errors could



**Figure 1.14:** With electric linear motors the wafer stage became “cleanroom-friendly”.

be reduced to a level of approximately  $1 \mu\text{m}$ , which was mainly limited by backlash, dynamics and friction. Experiments with piezoelectric actuators, to correct the remaining error, were not successful yet at that time, so the only solution that remained was to adapt the direct control principle of the VLP pick-up unit.

It is interesting to be aware of the following observation related to this step. The frame of thought of the traditional mechanical engineer was (and often still is!) based on the understanding that for precision something has to be stiffly connected on an adjustment mechanism that by itself is deterministically constrained in position. With the optical pick-up unit the approach was however completely different. In principle the optical pick-up unit is not connected to anything, while the position is determined by exerting electromagnetic forces that are controlled by the difference between the actual and the desired position, thus creating a virtual active stiff connection between the moving part and the track. This more abstract understanding that an accurate position can be realised with virtual stiffness appeared not easy to understand for people that are accustomed to realise more rigid mounted objects. Possibly for that reason only one real mechanical engineer was allowed in the optical research group at the Natlab and one of the authors of this book was one of those! The early breakthroughs in mechatronics have

been realised mainly by physicists and electronic engineers. An interesting observation!

Through the application of the principles from the VLP research in the so much larger system of the wafer stage it became possible to design a superior electric alternative for the hydraulic motors that took away one of the largest obstacles for customer acceptance of the first Philips wafer steppers. For this reason this electrical linear motor driven wafer stage, as shown in Figure 1.14, is symbolic for the success of the mechatronic discipline in the Netherlands.

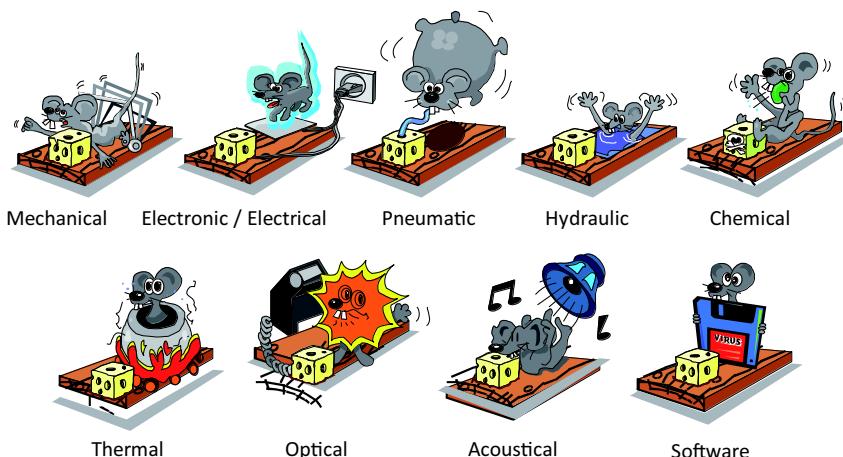
### **1.1.3 The impact of mechatronics on our world**

The described two subjects of research, in their mutual mechatronic relation within the same research group, have resulted in two significant technological developments with an influence far beyond the borders of the Netherlands. A world without optical registration of data, images and sound has become almost unimaginable and the Philips wafer stepper has grown within the Philips daughter ASML to the machine that determined the necessary conditions for the overwhelming developments in the electronic industry. The wafer stepper and its successor, the wafer scanner will be further presented in Chapter 9 at the end of the book where all the theory on mechatronics comes together.

## 1.2 Definition and international positioning

Mechatronic design is a discipline within mechanical engineering that combines the classical mechanical disciplines of static and dynamic mechanics, thermodynamics, metrology and tribology with historically non-mechanical disciplines like electronics, software and optics. This combination is certainly more than just the sum of these different disciplines. First and foremost, mechatronics should be considered to represent **technology integration to achieve optimal system functionality**. In fact a complete mechatronic product can only achieve its desired functionality through a process of systematic integration of all inherent disciplines, right through from the conceptual phase. Mechatronics opens up enormous technological possibilities, as already evidenced in the previous section by the appearance of sophisticated products like wafer scanners and compact disc players. These products would never have been realised by only a traditional single disciplinary approach.

Mechatronics is hence a real multidisciplinary field of expertise, a fundamental way of realising a certain specific function where in most cases controlled motion is determinative for the result. Being a design discipline, mechatronics focuses on integration and synthesis rather than analysis, though the latter is indispensable for a good understanding. Like shown in a humorist way in Figure 1.15, mechatronics aims to give a more optimal functionality than would be feasible by only focusing on one specialisation.



**Figure 1.15:** Essentially mono disciplinary solutions to a problem. They work, but is the solution optimal?

### **1.2.1 Different views on mechatronics**

Many definitions exist for the field of mechatronics. The first example finds its origin in Japan where the term mechatronics was first officially deposited:

**The planned application and efficient integration of**

- **mechanical and electronic technology**
- in a **multi-disciplinary and integrated approach of**
- **product and process design**

**to optimise production.**

This clearly is more focused on production rather than product development although it is both targeting application and integration.

From Europe comes the following definition which is closer to our understanding:

**A synergistic combination of**

- **precision mechanical engineering**
- **electronic control**
- **systems thinking**

**in the design of products and processes.**

This definition is however very wide and covers about every engineering subject ranging from controlling chemical processes to air planes. The relation with precision engineering is clear but precision is a relative concept. Present modern industries are all working towards the maximum of their technological capabilities. It is probably true to say that these definitions in fact define mechatronics to be the modern version of mechanical engineering. Mechanical engineering has in its history always integrated new emerging technologies, like for instance thermodynamics and electrical power sources. While the term “mechatronics” originated in the seventies of the last century, it clearly appears to have become an inextricable part of the present mechanical engineering discipline.

#### **1.2.1.1 Cultural differences in mechatronics**

The above definitions and the international developments in the field unveil some interesting observations regarding the differences in approach on

mechatronics. The understanding of these differences is important to be able to effectively cooperate with people from different regions and cultures.

First of all the term *Precision Engineering* originally was used for precision manufacturing and machining. Research and improvements in that field concentrated on primary processes like milling, grinding and polishing combined with extensive dimensional and geometric metrology to check and optimize the machining process. When electronics and control principles came available, this background has clearly resulted in the first mentioned Japanese definition of mechatronics. One might say that this definition is related to the *machining* orientation of mechatronics where the new technology was applied to correct flaws in the uncontrolled machines of that time. A clear example is found in the use of piezoelectric actuators to correct positioning errors in machine tools, where most of the linear movement was done by means of rotating motors and precision screw spindles. Without changing the basic concept for the *long-stroke* spindle positioning system, the piezoelectric actuator had its own feedback control system where it had to correct errors that originated from the screw-spindle actuator. This approach is still a practised way to achieve precision in machining tools with some clear benefits as disturbing cutting forces are absorbed by the inherently stiff construction of the piezoelectric actuator even before the feedback-control needs to correct them.

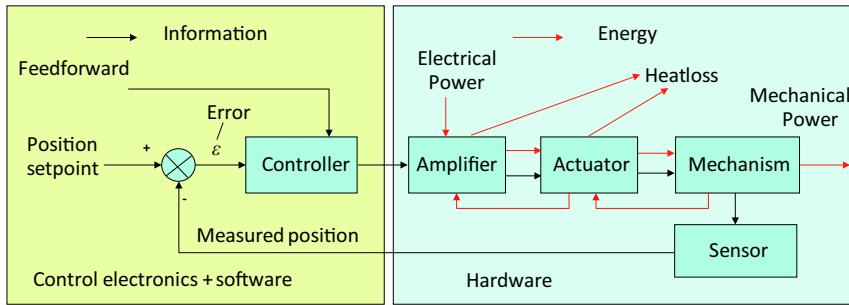
Meanwhile, in the essentially electronics and photon-physics oriented industry on consumer electronics and IC lithography exposure equipment, a different approach was followed. In the previous section it was shown how the development of contact-less optical recording with the use of optics had forced people in finding ways to manipulate photons in following a fast moving track. As photons are essentially without any mass or disturbing forces, it became preferred to avoid all contact with the “dirty” interfering vibrating environment and only control the system in respect to the track on the optical disk that had to be read out. This design concept resulted in a moving-coil actuated system with *zero-stiffness* to the surroundings and an electronic control system with inherent high virtual stiffness to the track on the optical disk. The resulting precision was impressive for that time ( $< 0.1 \mu\text{m}$ ). It had only one drawback because the possible movement range of a Lorentz actuator is limited or high cost of big magnets and coils would incur. This was eventually solved by adding a non critical long-stroke mover, the described screw spindle in the later CD players that only had to transport the stationary part of the moving-coil actuators. The demands on the long-stroke actuator were so limited that the ultimate cost of the total

system became much lower than would be the case with using piezoelectric actuators, as would have been the common approach in the machining industry.

The signatures of these two different applications are still very visible in the present mechatronic playground. In the USA, where the competition in the development and manufacturing of consumer electronic products is almost completely vanished, the machining industry determined the mechatronic research and development. Only in the last several years fast tool servo systems with reduced mounting stiffness and high bandwidth for increased precision are introduced, specifically for the manufacturing of high quality optical parts.

In Japan a mixture of both approaches is practised, as in that country both precision machining and consumer electronics are developed. Nevertheless the more conservative machining approach is frequently observed which sometimes hampers progress in their developments.

Europe is in this respect interesting in its own way. The largest part of Europe followed along the lines of the developments in the USA and focused on precision machining and related mechatronics instead of consumer products related technology. The Netherlands however went 100 % into the other direction. First of all there is hardly any machining industry left, while several world class companies in the high-tech industry are active in that country applying precision mechatronics in professional printing, in electron optic and x-ray imaging and in optical lithographic equipment. The related photon-physics oriented mechatronic approach has been instrumental in the success of these industries and for that reason it is unavoidable that this photon-physics oriented *Dutch school of mechatronics* resonates in this book.



**Figure 1.16:** A mechatronic motion system consists of several elements that are all required for the total functionality. The left side (yellow) only deals with information. Although fully analogue electronic realisations are still used, presently this function is mostly implemented in software with digital electronics. At the hardware side (blue) both information and energy is manipulated to achieve the final goal.

### 1.2.1.2 Focus on precision-controlled motion

In view of the aforementioned considerations it was decided to use the European definition of mechatronics and focus the contents of this book on controlled motion systems with emphasis on the precision engineering and positioning principles that are applied in the Dutch high-tech industry. A mechatronic system within this focus area generally consists of the following main components:

- A movable mechanical construction.
- An electrically controllable drive to move the mechanism in its degrees of freedom, the actuator.
- An amplifier to convert the electrical power to the power needed by the actuator as function of the control output.
- A measurement system to monitor the movement, the sensor.
- A control part that controls the system, the software.

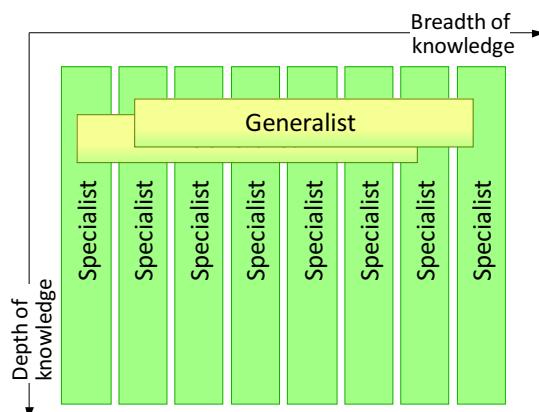
Figure 1.16 shows these elements in an overview. The arrows indicate the flow of information and energy, most of which takes place in the electrical domain. This connecting function of electricity within mechatronics is also one of the connecting elements in this book.

### 1.3 Systems engineering and design

Mechatronic design is very much related to the field of systems engineering. In this section the basic principles of systems engineering and design are introduced, because they are widely used in the high-tech industry and present a useful framework for the multidisciplinary design of any mechatronic system.

Systems engineering originated around World War 2 as a discipline to cope with the continuous increase in complexity of machines. Before that time, machines like airplanes, cars and ships were so fundamentally simple from a concept point of view, that almost without any exception several people knew how the total system worked. With the introduction of electronics, software and all kind of safety, diagnostics and other surrounding information oriented technologies in these machines, it became a hazard or even impossible to grab the whole entity by one person. This development has created the need to divide the capabilities in design between generalists, that have rather a global overview of the functionality of a system, and specialists, who have an excellent view on all details of a certain discipline. This is visualised in Figure 1.17.

In view of the multidisciplinary character of mechatronics, a mechatronic designer is rather a generalist than a specialist and he or she needs to keep a complete overview of a complicated system in order to be able to act as the lead-designer of a development team. That need has become more and



**Figure 1.17:** Both specialists and generalists are required to design a complex system.

(Courtesy of G. Muller, Gaudi site)

more a real issue as the capability to act as a real generalist with sufficient knowledge to effectively communicate with specialists in a multitude of disciplines is something one can hardly learn from books, but first and for all by experience.

Based on this trend, a continuous effort is taking place to find methods that help to keep the problems under control that are inherent to this increasing complexity. In the following, some basic methods are described that are widely accepted with clear benefits for a successful completion of a complex system design.

### 1.3.0.3 Definitions and V-model

As mentioned in the previous part, a consequence of the increased complexity of systems was the rise of specialists, people who knew a lot but only about a part of the system. With as example a car, one specialist knew all about the engine, one was specialised in the steering, one in the brakes, one in the suspension, and later even one for the electronics and software. One might think that it would be a perfect solution if these specialists just communicated with each other and by virtue of some overlap would come with an optimally designed car. Like with the mousetrap example this does not work in real life, unfortunately. Some people are more dominant than others and these often force a design in the direction of their hobby (read: specialism), which is not necessarily the optimal direction. Of course there are positive exceptions, but in many large companies and especially in situations where safety is an issue, methods had to be developed that would make the designs less depending on individuals, in order to achieve continuity in technological quality and performance. It was first in the professional, military and aerospace industry, that the concept of Systems Engineering was set forth. This concept is based on a large set of clear definitions and a visualised model, the *V-model of systems engineering*.

The following definitions originate with a few small adaptations from the International Council of Systems Engineering (INCOSE<sup>1</sup>) and the Gaudi website<sup>2</sup>, that provides systems architecture information and is maintained by Gerrit Muller, professor in Systems Engineering from the Buskerud University college.

---

<sup>1</sup>[www.incose.org](http://www.incose.org)

<sup>2</sup>[www.gaudisite.nl](http://www.gaudisite.nl)

**System:** An interacting combination of elements to accomplish a defined objective. These include hardware, software, firmware, people, information, techniques, facilities, services, and other support elements.

**Systems engineering:** An interdisciplinary methodology to enable the realisation of successful systems.

**Systems engineer:** An engineer trained and experienced in the field of systems engineering.

**Systems engineering processes:** A logical, systematic set of process steps selectively used to accomplish systems engineering tasks.

**Product Creation/Generation Process (PCP/PGP):** A logical, systematic set of process steps to create/generate a product or system.

**System architecture:** The arrangement of elements and subsystems and the allocation of functions to them to meet system requirements.

**System design:** The activity where it is determined how the system will be realised.

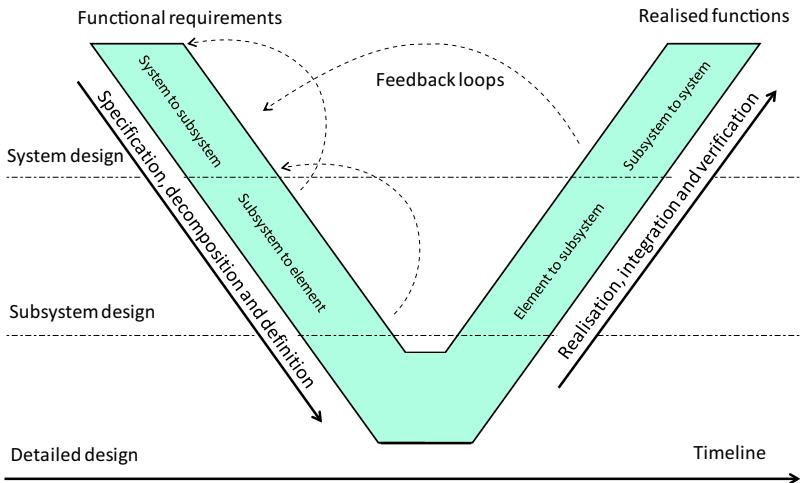
**System requirements:** The list of functionalities that the customer would like to have in the system.

**System specification:** The list of *unambiguously measurable* functionalities that are planned to be realised in the system.

**System performance:** The list of unambiguously measurable functionalities that are realised after completion of the system.

**System property:** The list of both wanted and unwanted both unambiguously measurable and not well measurable functionalities of a system.

It is important to note that in these definitions systems engineering is described as a methodology and approach to **organise** the design of a complex system. This means that the system engineer role in organising the process is not the same as the role of a system designer, who concentrates more on the content. In principle and even preferably these roles are united in one person.



**Figure 1.18:** The V-model of systems engineering divides the design process of a complex system both in time and in levels of complexity, in order to organise the effort over specialists and generalists.

The V-model of systems engineering as shown in Figure 1.18 was introduced in the design process to organise the way of working. The value of the model is found in the methodology of dividing responsibilities and tasks of both specialists and generalists over the complexity of the system. By working with layers with increasing detail per layer, the design of a complex system becomes manageable. Going down on the left leg of the V the following order of events is followed:

- In the “System design” top-layer, the functional requirements of the total product are determined and translated into system specifications and subsystem requirements.
- In the “Subsystem design” layer, these subsystem requirements are translated into subsystem specifications and requirements for the detailed elements.
- In the “Detailed design” layer, the requirements for the detailed elements are translated into drawings, software code and realised parts with a verified performance that can be assembled into functional subsystems.

Then the right leg is followed upward again.

- In the “Subsystem design” layer, the engineers integrate the different parts into a working subsystem that is tested in order to verify that the performance of the subsystems meets their specifications.
- In the “System design”, layer the different subsystems are integrated into a whole system and the performance of the total system is tested and verified.

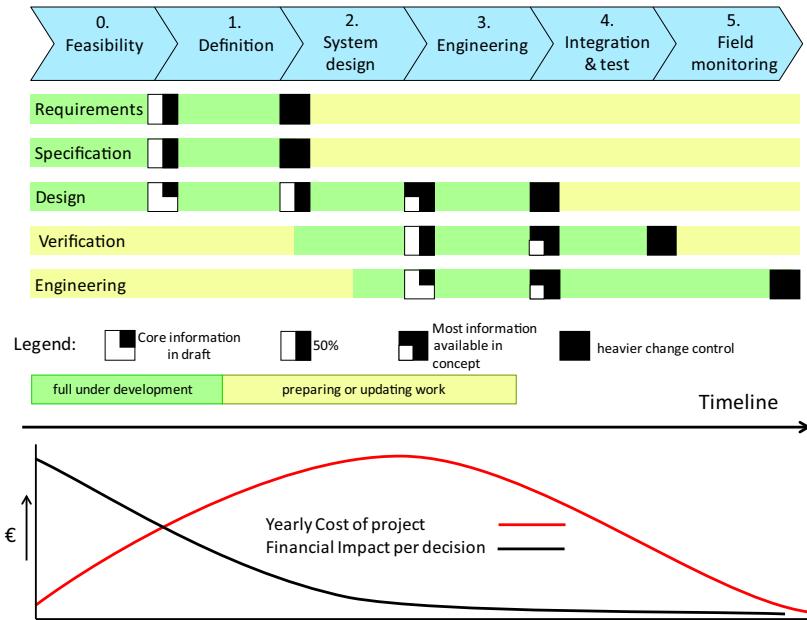
In very complex systems, the subsystem layer itself often consists of different layers. Each has its own requirements, specifications and realised functionality.

The work in each layer has a different character. The deeper one goes towards the bottom of the V-model the more detailed and concrete the work will be. But also within one layer the work is different between the left and the right leg, due to the fact that definition and realisation require different skills. For this reason, often the people in the team are changed during the project. This is not preferable from a learning perspective, because experience with results of the realisation phase will improve the decisions taken in the definition phase of a next project. Nevertheless, the work in the left leg is in practice done primarily by *innovator and creator* types of designers, while the right leg is realised by more *completer and finisher* types of people. In any case, frequent feedback is necessary within a project to overcome non compliances with the specifications that show up in the verification stage. Communication is also necessary between projects where experience from the past is used for continuous improvement over time. These feedback loops over different people need continuous attention as they do not happen automatically.

The subsystem layer is the place where many mechatronic systems are designed, like motion stages and actively controlled optical imaging systems. In the following section the work in these layers is described in the context of a regular development according to the *product creation process*.

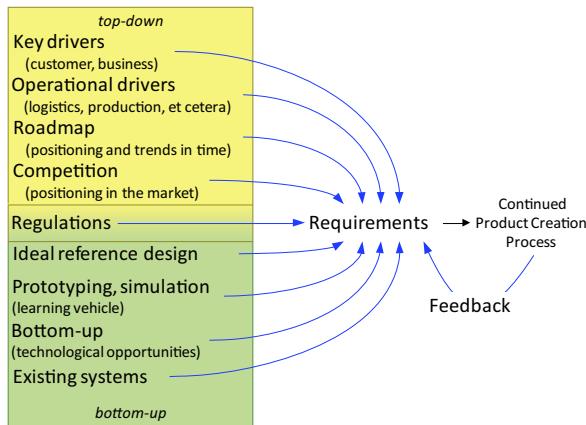
#### 1.3.0.4 The product creation process

Figure 1.19 shows the time division of the product creation process that is commonly used in industry. The phases 0, 1 and 2 correspond with the left leg of the V-model while the phases 3, 4 and 5 correspond with the right leg. From practical experience it is known that many mistakes are made during the feasibility and definition phase, when not sufficient time and effort is given to determine all factors that are at stake. In this early phase the



**Figure 1.19:** This standard time line of a product creation process is closely related to the V-model and is more directly suitable for planning purposes. The lower graph shows the yearly cost related to the effect of decisions on the financial results of the project and emphasises the need to especially control the first phases.  
(Courtesy of G. Muller, Gaudi site)

decisions with often the largest impact on the success of a project are made at the lowest relative cost. This relation is schematically shown in the lower graph of the figure. Especially in this phase the *systems engineer* needs to have a good overview on market demands and technological possibilities. In most cases he closely cooperates with a *product manager* who takes care of marketing issues and represents the end user of the product. This combination of often only two people is to a large extent responsible for the success or failure of a product. One might say that a badly started product creation process can hardly be repaired while of course a good start can still end up in a mess due to a bad execution. This duo of systems engineer and product manager has to deal with the tension from the external market in timing, product margins and functional requirements. They will define the final set of requirements in a process as shown in Figure 1.20 that are used as input for the development activities. They also have to keep these requirements closely under change control. After the integration and test of



**Figure 1.20:** Requirements originate from a large amount of input information. Well defined requirements are important for a successful end result.  
(Courtesy of G. Muller, Gaudi site)

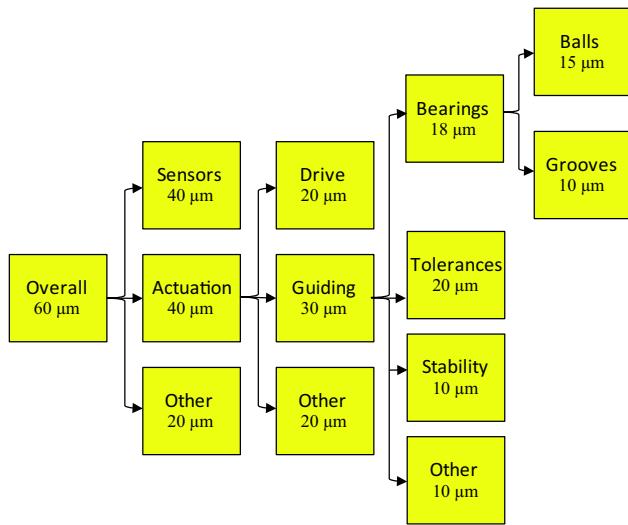
the product, this team is often also responsible for dealing with the market response on the product.

At all levels in the V-model, a trade-off has to be made between specifications of subsystems that have to work together. In the next section the related process of *requirement budgeting* will be explained a bit further.

### 1.3.0.5 Requirement budgeting

In Figure 1.21 an accuracy budget of a positioning system is shown as an example, indicating how the requirements and specifications are divided over the different subsystems with their corresponding sources of errors. It is clear that a systems engineer who has to define these requirements needs to have a thorough insight in all subsystems that contribute to the overall end result in order to give the largest error budget to the most difficult subsystem. The shown calculation is only valid if the errors are random stochastic and independent in which case one can take the square root from the sum of squares of all contributing errors as will be further explained in Chapter 8 on measurement.

The system engineer also has to decide which part of the budget is allowed to the smaller frequently occurring random error sources. This is noted in the item "Other" in the figure. This process of budgeting of requirements is only one half of the total activity as the top-down process to split-up the requirements is followed by a bottom-up process to determine the corre-

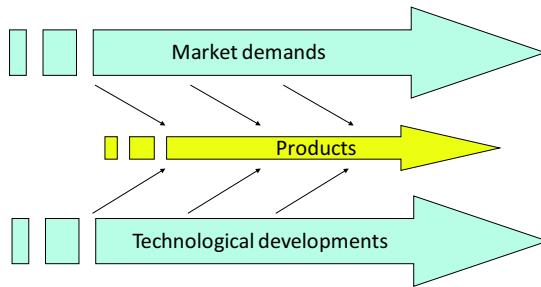


**Figure 1.21:** Even a simplified accuracy budget of a positioning system with many independent error sources already shows the choices that have to be made in close cooperation with the design teams that have to realise the targeted budget.

sponding specifications. The designers that are involved in this process will often face contradictory or conflicting requirements and in negotiation with the system engineer the specifications are determined with as goal a minimal effect on the original targeted full-system specifications. In this way the expertise of the responsible subsystem designers is used to create a balanced specification budget.

### 1.3.0.6 Roadmapping

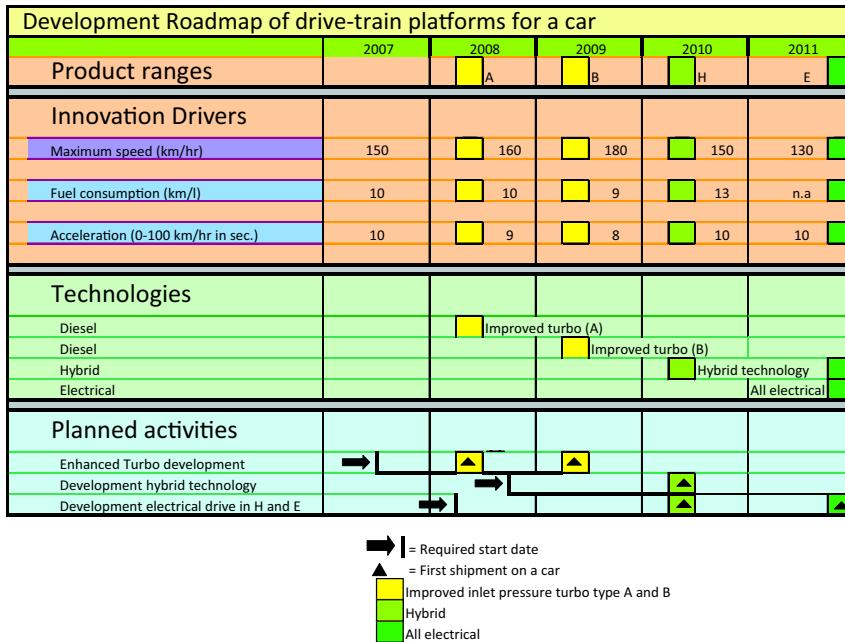
Figure 1.22 shows in a graphical way, that the process of developing products takes place in a continuous flow of changing external influences, the increasing demands of the market together with the also increasing technological developments. In most cases these external influences can hardly be influenced by the design team. A good example is the development of the Personal Computer. Initially the combination of IBM, Intel and Microsoft could determine their market approach without the need to take the competition into account. Inevitably however, at some moment in time, the competition like Apple and the Unix based Linux community had arisen forcing the seemingly monopolistic PC league to listen to their customers. This observation has as consequence that in general a lot of marketing research



**Figure 1.22:** The natural flow of product development.

is needed to get and keep a good feeling of the market dynamics and to make reliable estimates of future needs and possibilities. These are key tasks of the marketing people in close cooperation with their technological counterparts. For several practical reasons, developing products can never be as continuous as the market would like. To keep the cost down and the quality high, the diversity in parts and products has to be limited. For this reason products are generally developed in product families, also called *platforms*, that will stay in production as long as possible. As a consequence, when the time needed for the development itself is added, the initial decisions on requirements and specifications need to remain valid for a long time. One of the tools to systematically support these long term decisions is the process of *road mapping*. This is a strategic process that should be done at regular intervals to act as a basis for investments in new technologies and developments. Figure 1.23 shows a typical example of such a roadmap just for illustration of the principle. It is a highly simplified roadmap dealing with the engine drive-train of a car.

Of course there is a large freedom in methods of road mapping and the shown example is just one of the many possibilities to draw a suitable roadmap for the planning of future development activities. The roadmap consists of several layers where, like in the V-model, the upper layer deals with the total overview of all product introductions and the lower layer with the actual detailed plans on subsystem level. The definition of these roadmaps is real teamwork, where people from marketing, development, manufacturing and other sectors all need to have their input. An additional benefit of the process is the resulting commitment of all involved people to realise what has been agreed upon. It is in this way also clearly communicated and understood that some developments need more time than others, like the electrical drive in the example that is related to the applied batteries.



**Figure 1.23:** Roadmap that was drawn in 2007 as university lecture example, indicating the different engine-drive principles that were expected to be developed by the car industry in the following years.

A full roadmap of a product line can consist of many different roadmaps that are all interrelated. A suitable method is again based on the V-model with an overall roadmap on system level and more detailed roadmaps on subsystem level.

### 1.3.1 Design methodology

As noted in the previous part a (system) designer has to deal with several contradictory demands:

- Shorten the development time in order to get the product earlier in the market (more profit!).
- Stimulate the market with a continuous flow of new products.
- Enhance product performance by adding functionality (more cost!).
- Reduce product cost to increase margins (more profit!).

Those items need to be balanced as they can not be achieved simultaneously in a simple way.

A good example of this balancing is the observed phenomenon that, without additional measures, a reduction of the development time will almost by definition increase the total cost of the development process in an exponential way, while the gain in profit for an earlier introduction not always justifies such increased cost. Especially in the high-tech electronics industry, however, an earlier introduction can be very valuable. A good example of the first three bullets is the dramatic diversification in mobile phones and PDAs. Being often an impulse buy where the choice is highly determined by emotion. As a consequence of this emotive value, an old design is often obsolete within a few months. This problem is far less the case for large domestic appliances, like washing machines, that mainly have to be reliable with a very long lifetime without any surprises in their user interface.

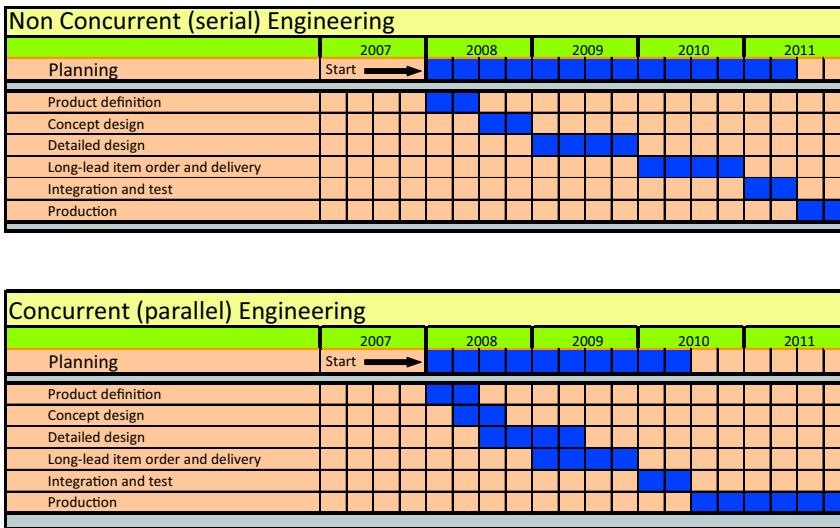
Because of these contradictory demands several methods are used of which the following two will be presented in the last part of this chapter:

- Concurrent Engineering.
- Modular design with platforms.

### 1.3.1.1 Concurrent engineering

The drive for *total quality* in the second half of the last century has forced development activities to become a more and more rigidly planned process. Total quality means that not only the products are expected to function exactly like they are promised to do, but also the process to create and produce them needs to be fully determined. One item that was high on the priority list at that time was the reliability of the development timing and the reduction of scrap due to unplanned repairs. This process resulted in a strict planning with milestones and go/no-go decisions in order not to start with a next step until all questions of the previous step were answered. This approach had several advantages:

- The planning became continuously more solid towards the end of the process.
- High investments could be postponed until really needed.
- Surprises were avoided by testing prototypes in the period that *long-lead items* were ordered.



**Figure 1.24:** Concurrent engineering results in shorter lead times by working in parallel on different phases of the project. It requires a high discipline in communication and change control.

- Manufacturing could better prepare itself.
- Development personnel could make the product and service documentation during the *order and delivery* phase of the parts.

The process to allocate all activities essentially in series over time often resulted in practice in a development lead time of more than three years for a general consumer product. To partly solve this problem, the process of concurrent engineering was introduced as shown in Figure 1.24. Under pressure of market needs, people started to make use of the fact that in each phase of the development process some elements will take more time than others to be finished. It might be useful to treat these long-lead items as *off-the-shelf* parts that have to be used just as they are. This approach means that the long-lead items should be the first to design and the other parts have to adapt to the interfaces chosen for these long-lead items. In this way a hierarchy of parts is created that enables the designers to start earlier with the next step in the process.

An important condition to make this method work is the use of agreed interfaces. From the computer industry several well established interface standards are known like PCI, USB and (serial-) ATA to name a few and these have helped this industry to become as successful as it is. For the

design of mechatronic systems the situation is less easy due to the low flexibility of the solid hardware but even in that field standard interfaces are quite common with as best example the standardisation in connective parts like bolts and nuts

In spite of the reduction of the development time, this more parallel concurrent engineering approach also has some drawbacks that need attention:

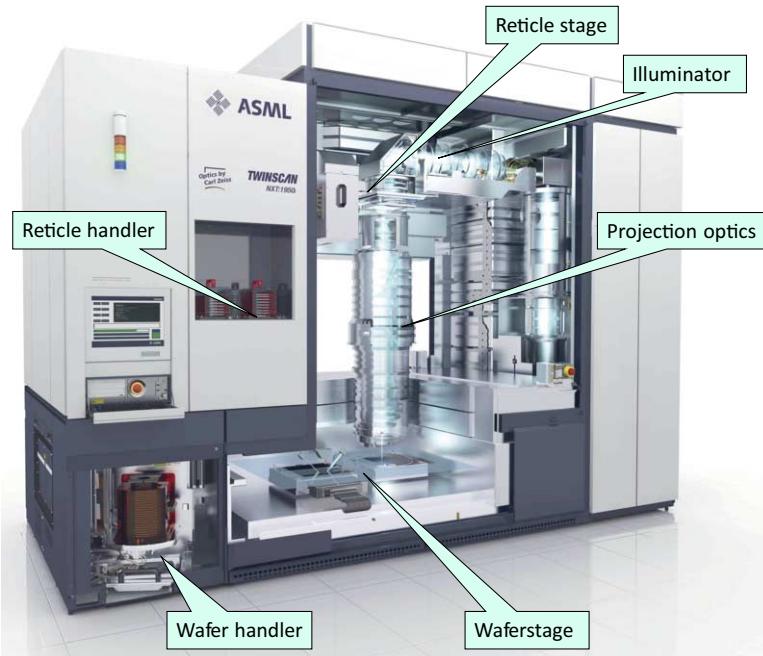
- Very costly repairs in case of forgotten or overlooked interface needs.
- Less time available for documenting.
- Less gradual planning of development resources.

In the industries where concurrent engineering became the “de facto” standard of operation like in semiconductor manufacturing equipment, several methods are used to avoid these mentioned problems. First of all a formal and intensive communication network is created, focused on exact interface definitions. Specifications are only adjusted in a disciplinary process according to the rules of systems engineering. Secondly it appeared to be of crucial importance to work with a modular design that is based on a technology platform with a lifetime of more than one product introduction.

### **1.3.1.2 Modular design and platforms**

The first example of a modular design around platforms is seen in the previously mentioned computer industry. A basic desktop PC consisted around the time of writing of this book of several clearly distinguishable modules, like the power supply, motherboard, memory, processor, hard disk, monitor, keyboard and several other smaller items. All of these are modules in principle interchangeable for other versions within a platform, where a platform is in this case a certain combination of processor and related hardware. The designers of the platform determine the interfaces that communicate with the peripheral electronic modules. This way of working has enabled many electronic companies to deliver parts with different levels of functionality. Together with the use of standardised hardware-connectors and dedicated driver-software, they were often even capable to provide solutions for different platforms.

Another example of a modular design is found in the mobile phone industry. By exchanging panels and windows on a standard electronic circuit, a large variety of products can be derived without changes on the basic functionality



**Figure 1.25:** The main modules in a wafer scanner that is based on the “Twinscan” platform of ASML. All modules are individually upgradable giving the platform a virtually infinite lifetime in a market that demands a continuous increase in performance.

(Courtesy of ASML)

and reliability in the new design. While the power supply and display are controlled with embedded software (firmware), much flexibility is available in showing or hiding functions depending on the market position of the specific product. Most functions, sometimes including the camera, are integrated and depending on the execution they are switched on or off in the firmware or just covered by the shell. The total cost of such a design is often even lower when some hardware in the product is not used, than would be the case if a special hardware version without that specific function has to be developed. This is mainly due to the lower logistic cost related to the reduced diversity.

The last example of a modular design is the ASML wafer scanner as shown in Figure 1.25. The “Twinscan” platform with a dual wafer stage became the leading technology in this industry. In retrospect, many reasons for this success can be mentioned like the team spirit, commitment and drive of the people, the partnership with Zeiss and the strong mechatronic roots in

Philips. Nevertheless the modular design approach proved to be one of the key elements of success for the following reasons:

- All modules can be developed on different locations.
- All modules can individually be tested to their full specifications on test rigs without the presence of the other modules.
- All modules can be mounted and dismounted from the machine without affecting the other modules.
- Duration of the installation times at the customer site is significantly reduced.
- Upgrading is possible even at the customer site, which makes most machines last almost forever.
- Testing and servicing is relatively easy because all important modules can be dismounted.

# **Chapter 2**

## **Electricity and frequency**

The performance of a mechatronic system is related to the ability to generate and control a certain movement of a body. In general this means that forces need to be applied to keep the body on track while compensating other, external forces. In case of constant external forces, there is no real difficulty to control these but in reality external forces are never really constant nor are the movements that need to be controlled. These forces and movements change in different ways both periodic and stochastic and this means that the performance must be analysed by looking at the response of the system to periodic signals of different frequencies. These frequencies are often observed in both the mechanical and electric domain.

This chapter begins with the principles of electricity followed by a presentation on the concept of frequency with periodic signals and wave propagation. The frequency and time domain related Fourier and Laplace transform methods are presented as necessary mathematical background that is used in the dynamic analysis of mechatronic systems. The chapter finishes with some frequently applied graphical representations of frequency and time responses of a dynamic system.

## 2.1 Electricity and signals

The phenomenon of electricity is fundamentally based on the electric charge composition of atoms. Atoms have a core, consisting of neutrons and protons, with a positive charge determined by the amount of protons, while neutrons are not charged. Around the core a “cloud” of electrons with a negative charge neutralises the charge of the core in a stable situation where the amount of electrons equals the amount of protons.

The unit of the amount of charge is [C] (Coulomb), named after the French physicist Charles-Augustin de Coulomb (1756 – 1806), because of his work on electricity and magnetism. As is true for many definitions in electricity, the value of one Coulomb has been determined much later than when the initial definitions were made. Electrons and protons were not known when the first electric experiments were carried out. It appeared that 1 C equals  $6.25 \times 10^{18}$  times the charge of one electron.

Electrons are not uniquely linked to their atoms but can more or less freely move. The freedom to move is inversely proportional to the electric *Resistivity* ( $\rho_r$ ) [ $\Omega\text{m}$ ] of the material. A material can be an insulator with a very high resistivity value like glass with  $\rho_r \approx 10^{10} \Omega\text{m}$  and even higher with plastics that might have a resistivity of  $\rho_r \approx 10^{20} \Omega\text{m}$ . If the resistivity is very low, the material becomes a conductor like Copper with  $\rho_r = 1.68 \cdot 10^{-8} \Omega\text{m}$ . Semiconductor materials are a special category having an electric resistivity in between a conductor and an insulator. Unique for these materials is the possibility to change their resistivity over a large range from fully isolating until highly conducting, either permanently by physical/chemical modification or dynamically in an electronic circuit.

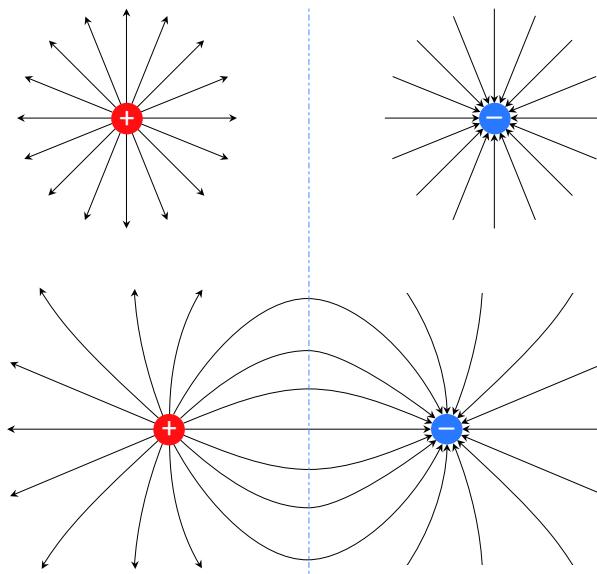
The resistance ( $R$ ) in Ohm [ $\Omega$ ] of an object with a cross section ( $A$ ) and a length ( $l$ ) is directly derived from the resistivity of the applied material and is equal to:

$$R = \rho_r \frac{l}{A} \quad [\Omega] \tag{2.1}$$

It is clear that the resistance increases proportional with length and resistivity and decreases with a increase of the cross section.

### 2.1.1 Electric field

Charged particles like electrons and protons have as property that they either attract each other, when their charge has an opposite sign, or repel



**Figure 2.1:** Charges have an electric field that is directed away from the positive charge and towards a negative charge. The electric field is graphically represented by field lines. The density of the field lines determines the force acting on other particles in the vicinity and the arrows of the field lines give the direction of the force in case of a positive inserted charge. The combined field of two opposite charges results in curved lines that start at the positive charge and end at the negative charge.

each other, when their charge has the same sign. For example an atom with a lower number of electrons than its proton count will attract electrons until its charge is neutralised again. The related *electrostatic force* is the driver behind all important electric phenomena. The magnitude and direction of this force is represented by the electric field  $\mathbf{E}$ . Mathematically the electric field at a distance  $r$  in the direction of unit vector  $\hat{\mathbf{r}}$ , with its origin in the centre of a single charged particle with charge  $q_x$ , is represented by the following equation:

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi r^2} \frac{q_x}{\epsilon_0} \hat{\mathbf{r}} \quad [\text{V/m}] \quad (2.2)$$

The term  $\epsilon_0$  is the *electric permittivity in vacuum* and is equal to a value of  $\epsilon_0 \approx 9 \cdot 10^{-12}$  [AsVm]. From the  $4\pi r^2$  term in the denominator it can be concluded that the magnitude of the electric field is related to the surface of a surrounding sphere at distance  $r$ . This can be visualised by means of *field lines* as shown in Figure 2.1. The field lines represent the direc-

tion of the electric field as arrows and their density is proportional to the magnitude of the field. This also means that the field is constant orthogonal to the field lines. By the definition of unit vector  $\hat{\mathbf{r}}$ , the direction is outward for a positive charge and inward for a negative charge. Though the drawing is two-dimensional, in reality the space around the charge is three-dimensional, which corresponds with the decrease in the electric field and the corresponding density of the field lines to the distance squared. In the lower drawing of the figure the effect on the course of the field lines of two opposite charges is shown. The connecting curved field lines are the result of the superposition principle of electrostatic fields, which means that fields can be simply added together. As a consequence, the resulting field at the equidistant plane from both charges is zero and the field lines will cross this plane perpendicular, leading to the shown curved lines from one charge to the other.

The magnitude and direction of the force that is acting on a charged particle with charge  $q_x$ , inserted in a field  $\mathbf{E}(\mathbf{r})$  that originates from several other charged particles, can be calculated quite straightforward as follows:

$$\mathbf{F}(\mathbf{r}) = q_x \mathbf{E}(\mathbf{r}) \quad [\text{N}] \quad (2.3)$$

This force  $\mathbf{F}$  will be directed in the direction of the field in case of a positive inserted charge and in the opposite direction in case of a negative inserted charge. This direction corresponds with the known phenomenon that negative charges are attracted in the direction of a positive charge.

### 2.1.1.1 Potential difference

The vectorial nature of the electric field and the related force is very useful when determining physical effects in for instance electron microscopes, where charged particles have to be controlled in velocity and direction. In electric systems, where electrons are guided in conductive material like resistors, wires and other elements, it is more easy to work with the scalar magnitude of the related potential energy of a charged particle that is inserted in an electric field. What happens can best be explained when looking at two points in the electric field at a different distance from the first charge. The potential energy is determined in the same way as with gravitation. When a particle with charge  $q_x$  moves from location  $\mathbf{r}_1$  to  $\mathbf{r}_2$ , the work exerted by the electrostatic force will change its potential energy. The energy ( $E$ ) is given in the unit Joule [ $\text{J}$ ] named after the English physicist James Prescott Joule (1818 – 1889) for his important exploratory work on

energy. The change in potential energy of a charged particle has a negative sign, when the movement is in the direction of the force and is given by:

$$\begin{aligned}\Delta E_p &= - \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{l} \\ &= -q_x \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{E} \cdot d\mathbf{l} \quad [\text{J}]\end{aligned}\tag{2.4}$$

Due to the sign convention this means that a negative charge will have a positive potential energy in the vicinity of a positive charge which increases with the distance. This reasoning corresponds with the potential energy of a mass under the gravity force on earth. From this difference in potential energy between location  $\mathbf{r}_1$  and  $\mathbf{r}_2$ , the *potential difference* can be defined in order to get a number that is independent of the charge of the inserted particle, by dividing the previous equation by  $q_x$ .

$$\begin{aligned}\Delta V &= \frac{\Delta E_p}{q_x} \\ &= - \int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{E} \cdot d\mathbf{l} \quad [\text{V}]\end{aligned}\tag{2.5}$$

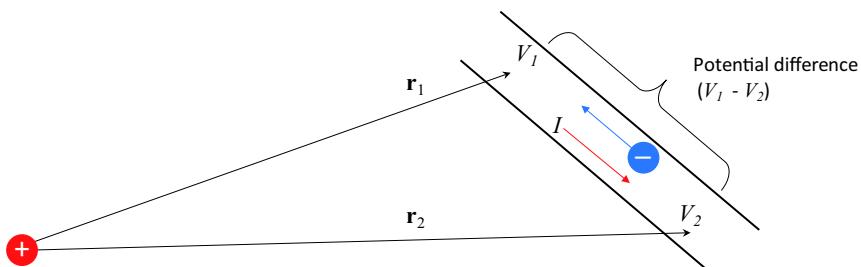
This potential difference  $\Delta V$ , or just shortly  $V$ , equals the electric potential between  $\mathbf{r}_1$  and  $\mathbf{r}_2$  and gives the electrostatic potential energy per unit of electric charge between these points.

The potential  $V$  reduces the vector field  $\mathbf{E}$  to a scalar field without losing information. The unit Volt [V] for the potential difference is named after the Italian physicist Count Alessandro Guiseppe Antonio Anastasia Volta (1745 – 1827), because of his work on the first electric batteries.

As with all forms of potential energy, this potential  $V$  is a relative quantity as it only gives a potential difference. An absolute value can be obtained by defining the potential at a certain point to be zero. In electronics this point is the common ground. In electrostatics often a point at infinite distance is used. This will give the following “absolute” potential and “absolute” energy level:

$$V(\mathbf{r}) = - \int_{\infty}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{l} \quad [\text{V}] \implies E_p(\mathbf{r}) = -q_x \int_{\infty}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{l} \quad [\text{J}]\tag{2.6}$$

These are positive numbers, when the  $\mathbf{E}$  field is created by a positive charge, as this field is directed outwards. These expressions make clear that the



**Figure 2.2:** Two points at a different distance from a charge have a difference in potential. When the charge is positive, the potential  $V_1$  is higher than the potential  $V_2$ , which means  $V_1 - V_2$  is positive. A negative charge, trapped in a conducting wire between  $\mathbf{r}_1$  and  $\mathbf{r}_2$ , will be driven by the electrostatic force in the direction of  $\mathbf{r}_1$ , which corresponds with a positive electric current flowing in the other direction.

potential difference is equal to the derivative of the potential energy over the charge. This is important in the next section when changes in charge represent a current.

### 2.1.1.2 Electric field in an electric element

A non-zero electric field determines a potential difference in space and vice versa. This means, that in situations where a potential difference is present, for instance over a resistor, an electric field is present inside the element. This field exerts forces on the charges in the resistor, causing the free charges to move. To visualise this, a thought experiment might help by assuming in Figure 2.2, that an electron can be “trapped” at position  $\mathbf{r}_2$  with potential  $V_2$  in a tunnel, that only leads to  $\mathbf{r}_1$  with potential  $V_1$ . In that case the electron will move to  $\mathbf{r}_1$ , because of its higher positive potential, when using Equation (2.6) referenced to infinity. If the potential had been calculated relative to the centre of the positive charge instead of infinity, so integrated from  $\mathbf{r}$  to zero, a negative potential would have resulted, increasing as function of the distance  $|\mathbf{r}|$ . In this thought experiment it is observed, that the electron seeks the position with the lowest negative potential energy, referenced to the positive charge. Indeed when  $|\mathbf{r}| = 0$ , the energy is lost as both opposite charges will cancel each other out. On the other hand, if this process would be reversed by taking an electron away from a location, this would result in a positive charge at that location and an increasing potential difference at increasing distance. It is this principle that creates the possibility to realise an electric circuit.

## 2.1.2 Electric current and voltage

Moving charges are known as electric current, which is given the variable-symbol  $I$  with unit Ampère [A] named after the French physicist and mathematician André-Marie Ampère, because of the high value of his research on electromagnetism. The current is equal to the flow of charge  $C$  per unit of time through a surface [C/s].

### 2.1.2.1 Voltage source

By actively separating charges, a source of potential difference is created, a *voltage source*. The name voltage for potential difference is more common in electronics, even though the name and the unit are mixed and the relative meaning is lost. A voltage source has two interface points with the outside world, called *electrodes* or *terminals*, of which the positive electrode has a more positive potential than the negative electrode. The voltage source can supply a continuous flow of electrons, running from the negative electrode towards the positive electrode through an external conductive load. This electron flow is represented in electric circuits by a positive current, because electrons are negatively charged, flowing in the other direction from the positive electrode to the negative electrode.

To achieve the separating action between the negatively charged electrons and the stationary positively charged protons, a voltage source can be imagined to possess an internal electric field, that is directed from the negative to the positive electrode. Like explained in the previous section, this field drives the electrons in the opposite direction, away from the protons towards the negative electrode. The resulting surplus of electrons at the negative electrode with the corresponding lack of electrons at the positive electrode results in a charge and potential difference between the electrodes, with a corresponding electric field outside the source. This external electric field is equal to the internal “driving” field, however it is pointing in the opposite direction.

Because of these opposite directed fields, the previous reasoning on the potential difference is confusing as it was defined as the negative integral of an electric field over a certain distance. This is still valid for the external field, with a positive potential difference between the positive electrode and the negative electrode. For the internal field this same potential difference should be calculated by taking the positive integral of this field. To avoid this confusion, the positive integral has been given a different name, the

### *Electromotive Force $\mathcal{F}_e$ .*

When  $\mathbf{E}_i$  is the internal electric field,  $e_1$  and  $e_2$  are the electrode locations and  $d\mathbf{l}$  is the inner path between the electrodes, the electromotive force becomes:

$$\mathcal{F}_e = \int_{e_1}^{e_2} \mathbf{E}_i \cdot d\mathbf{l} \quad [\text{V}] \quad (2.7)$$

In Chapter 5 an example of a voltage source will be presented based on induction of an electric field by a changing magnetic field. This is the most common method of generating electricity. This principle will show to be fully compatible with the mentioned integral expression. With the other well-known method to generate electricity, based on chemical processes in *batteries*, this is less straightforward to reason as the potential difference is the result of the difference in chemical potential of the applied elements. Especially with a battery it is more easy to just use the electromotive force as a given.

The value of this electromotive force is equal to the potential difference of the external field when no load is applied. In that case, due to the equilibrium between the electromotive force and the external potential difference, no further separation of charge can take place. This means that no current will flow inside the voltage source, when it is not connected to a load. As soon as an electrically conducting load is connected to the voltage source, the potential difference will drive electrons from the negative electrode through the load to the positive electrode. Due to the required balance between the electromotive force and the potential difference, the amount of electrons flowing through the load will be replenished by a flow of “new” electrons, driven by the electromotive force. This process of continuous electron flow is equivalent to a continuous flow of positive current, inside the source from the negative electrode to the positive electrode and through the load from the positive to the negative electrode.

This reasoning is valid for an “ideal” voltage source. In reality a voltage source will also show some internal imperfections that cause the voltage at the electrodes to be different from the electromotive force, depending on the current level. This phenomenon will be presented further in Chapter 6 on electronics. In that chapter also the *current source* will be introduced, that drives an electric current through a load with a magnitude, that is independent of the load. It has an infinite capability to adapt the electromotive force to any load, such that the current remains constant.

The following summarises the most important relations:

- An electric current is a physical representation of a positive entity. The real current in most electric circuits<sup>1</sup> is the movement of the electrons in the opposite direction of the current due to their negative charge.
- The electric field inside a source or a load is pointed in the direction of the positive current.
- A voltage source is based on a sustained electric field, represented by its electromotive force, that drives the electrons to the negative electrode.
- The voltage source creates an external electric field, represented by its potential difference, that is directed from the positive to the negative electrode, so opposite to the internal electric field in the source. the external potential difference drives a current through an electric load from the positive to the negative electrode.
- The electromotive force from the source is equal to the potential difference of the external electric field and is mostly just called “voltage”.

### 2.1.2.2 Electric power

When a charge moves through an electric field, its potential energy  $E_p$  and its electric potential  $V$  changes. The power ( $P$ ) of a system is defined as the time derivative of energy with unit Watt [W], named after the Scottish mechanical engineer James Watt, because of his famous work on steam engines. With this definition the power can be expressed as follows:

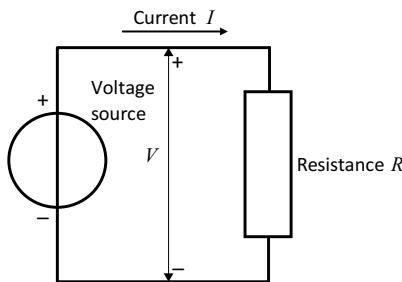
$$P = \frac{dE_p}{dt} \quad [\text{W}] \quad (2.8)$$

The electric power can be expressed in the electric potential of the charges as given in Equation (2.5), and in the current:  $I = dq/dt$  [A]. This gives the familiar expression for electric power:

$$\begin{aligned} P &= \frac{dE_p}{dq} \cdot \frac{dq}{dt} \\ &= V \cdot I \quad [\text{W}] \end{aligned} \quad (2.9)$$

---

<sup>1</sup>In solid material positive charge can not move. The only examples of a material flow corresponding to a positive current is in ionised gases and the movement of positrons but the latter do not play a role in the real engineering world.



**Figure 2.3:** Basic electric circuit consisting of a voltage source and a resistive load.

The current is proportional to the voltage and inversely proportional to the resistance, following Ohm's law. The notation of the potential difference is given with a double tip arrow where the signs are noted to define the direction from + to -.

The signs of the previous equations are very important. It was shown, that an electron as being a negative charge moves towards a decreasing negative potential while a positive charge on its turn would experience a force in the direction of a decreasing positive potential. This means that the electric potential energy of a charge decreases due to this movement and the corresponding power becomes available in another form, for instance the dissipated heat in a resistor or in mechanical work in an electro motor. In this book the positive direction of a current is defined by the pointing direction of a single arrow while the positive direction of the potential difference is given by a + and - sign adjacent to a double arrow to indicate the relative potential difference.

It should also be noted that the units after the equations will most often be omitted when it is obvious what is meant.

### 2.1.2.3 Ohm's law

Figure 2.3 shows the most basic electric circuit possible. It consists of a voltage source and a load, connected between the electrodes, that pulls current from the source. In this basic form of an electric circuit the load is a resistor  $R$  with unit Ohm [ $\Omega$ ]. In the next section the load will be extended into a complex load, but originally the German physicist Georg Simon Ohm (1789 – 1854) based his well-known law on the research he did using the electric batteries of Volta. With the constant, not time dependant, voltage of batteries the current in the circuit is only dependent on the resistivity in the electric circuit. For this reason Ohm's law gives the relation between

the current  $I$ , voltage  $V$  and resistance  $R$  of this electric system:

$$V = IR \Rightarrow I = \frac{V}{R} \quad (2.10)$$

With this clear relation the electric power delivered by the source, and dissipated as heat in the resistance, is as follows:

$$P = IV = I^2R = \frac{V^2}{R} \quad (2.11)$$

This squared, non-linear relation is an important phenomenon in mechatronic systems. As an example of this phenomenon, an increase of the force of an actuator requires an increase of the current. This increase results in a squared increase of the power dissipated in the series resistance of the actuator. A configuration that was just safe from a thermal point of view, can become suddenly overheated, when only a seemingly limited increase of the current is applied.

#### 2.1.2.4 Practical values and summary

Practical values of voltage show a very wide range:

- MV : The level of a lightning stroke.
- KV : The level used in power distribution.
- V : Mains supply, power supply of electronics, digital electronics.
- mV : Small signals in measurement sensors.
- $\mu$ V : Extremely small signals, “buried” in noise and interference.
- nV : Unmeasurable.

In practice the term nV is only found in expressions like  $nV/\sqrt{\text{Hz}}$ , that relates to the density of noise in a certain frequency band.

Practical values of current show an even wider range:

- kA : Power distribution.
- A : Mains, household equipment, actuator input current.
- mA : Small signals, power supply of low power electronic circuits.
- $\mu$ A : Small signals, sensors.

- nA : Extremely small signals, “buried” in noise and interference.
- pA : Hardly measurable, Scanning Tunnelling Microscope (STM) probe.
- fA : Unmeasurable.

In table 2.1 the most relevant definitions are summarised.

**Table 2.1:** SI units for electricity.

Physical quantity	SI unit	Variable	Relation
Charge	Coulomb [C]	$C$	
Current	Ampère [A]	$I$	$A = C/s$
Potential difference	Volt [V]	$V$	$V = J/C = Nm/As$
Resistance	Ohm [ $\Omega$ ]	$R$	$\Omega = V/A$
Power	Watt [W]	$P$	$W = VA$

### 2.1.3 Variability of electric signals

Electric voltages and currents are generally not constant over time. The variability of these voltages or currents contain information about their origin, for which reason they are called *electric signals*.

In physics a signal represents any value with a variability over time and/or space, the *temporal or spatial variability*. An example of temporal variability is sound that changes over time in amplitude and tonal character. In mechanical engineering a position dependent force is an example of spatial variability. In optics, spatial variability represents the difference in intensity of light as function of the location on a surface. In a mathematical sense both spatial ( $f(x)$ ) and temporal ( $f(t)$ ) variability can be treated the same way, when  $x$  is exchanged with  $t$  or vice versa in the relevant formulas.

Electric signals in mechatronic systems are always temporal even though the signal can be derived both from temporal and spatial physics phenomena. The latter are translated into the time domain by for instance a scanning movement with a constant velocity and measuring with a clear time reference. For this reason, the following part of this chapter will mainly deal with temporal varying signals while in Chapter 7 on optics the spatial variability will mainly be used.

### 2.1.3.1 The concept of frequency

The variation of an electric signal over time can be seen as a combination of three types of behaviour:

- Constant, unidirectional, called *DC* from Direct Current.
- Periodic alternating, bidirectional, called *AC* from Alternating Current.
- Random, stochastic, called *noise*.

The DC value of an electric signal equals the average value of the current over time. For an alternating signal the term *temporal frequency* is defined as the number of occurrences of a certain event per unit of time. Consequently one can define the time period related to this frequency as the time taken for one event to happen. In SI units the frequency with variable-symbol ( $f$ ) is expressed in the unit Hertz [Hz] named after the German physicist Heinrich Hertz (1857 – 1894). The time period of one cycle, with variable-symbol ( $T$ ) in seconds [s] is inversely proportional to  $f$ .

$$T = \frac{1}{f} \quad [\text{s}] \quad (2.12)$$

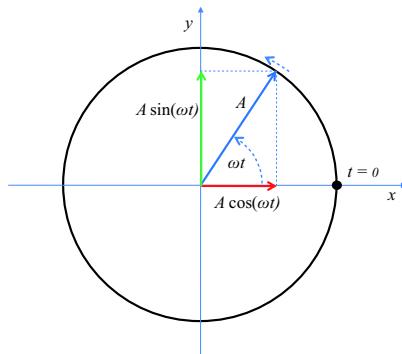
In physics and engineering the term *angular frequency* with variable-symbol ( $\omega$ ) is often used. This frequency is directly related to the mathematical description of harmonic oscillations, defined by the sine or cosine of an angle changing with a constant angular velocity  $\omega$  and an amplitude  $A$ , as shown in Figure 2.4

$$x(t) = A \cos(\omega t) \quad y(t) = A \sin(\omega t) \quad (2.13)$$

When the sine and cosine function are presented in a graph as function of the angle, this results in a graphic wave shape as shown in Figure 2.5. Though this is not a real physical wave, the term “waveform” is often also used for this graphic representation of a frequency because of its shape. When  $\omega$  is constant, the horizontal axis is also a time line where  $2\pi$  equals the period  $T$ . This means that the angular frequency is directly related to the temporal frequency, because one period  $T$  equals a full circle being an angle ( $\phi$ ) of  $2\pi$  radians:

$$\omega = 2\pi f \quad [\text{rad/s}] \quad (2.14)$$

It is very important to be aware of these different units as this is one of the traps even experienced designers encounter. A factor 6 is almost an order

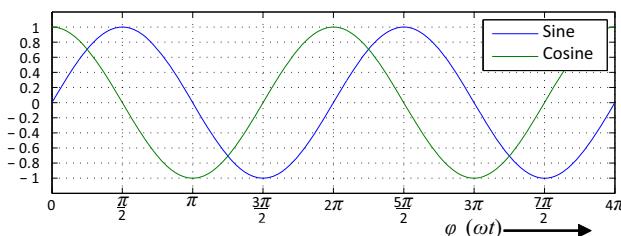


**Figure 2.4:** The sine and cosine of angle  $\phi = \omega t$  represented in a plane.

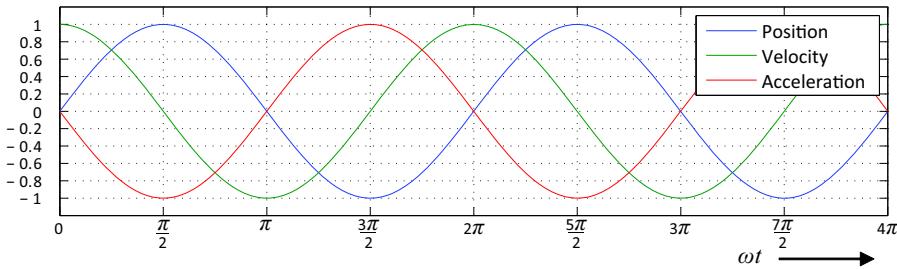
of magnitude and can easily result in wrong conclusions. In this book both units will be used. The [Hz] units are applied with the practical examples and graphical representation, where the relation with periodicity and multiplicity of events is more clear. It is often more related to the real world and human understanding. The [rad/s] unit is mostly used for mathematical analysis, equations of motion and the relation with physical effects.

When two signals with the same frequency like the sine and cosine of Figure 2.5 are compared, than the sine is shifted with an angle  $\phi = \pi/2$  in respect to the cosine. This time relationship is called *phase* and is one of the most important parameters in a controlled motion system because a negative phase shift represents time delay, that can cause instability when applying negative feedback. For practical reasons this phase shift is expressed in degrees ( $^\circ$ ), because of the small numbers involved. A  $30^\circ$  phase shift can have a large impact while it is less than  $\pi/4$ . Though the degree is not an official SI unit, it is widely accepted to be used in this context.

A simple thought experiment can be done on position, velocity and acceleration with a body that moves with a sinusoidal movement  $x(t) = \hat{x} \sin(\omega t) =$



**Figure 2.5:** Sine and cosine function of angle  $\phi = \omega t$ .



**Figure 2.6:** Two periods of the position, velocity and acceleration of a body with a sinusoidal movement:  $x(\omega t) = \hat{x} \sin(\omega t)$  with  $\hat{x} = 1$  and  $\omega = 1$ .

$\hat{x} \sin(2\pi f t)$  with an amplitude  $\hat{x}$  of 1. The blue line in Figure 2.6 shows that the position at  $\pi/2$  is stationary. This means that the velocity is zero, corresponding to the green line which was drawn at the cosine function. Also when observing for instance the steepest down slope of the position at  $\pi$  it shows that then the velocity is maximum negative, which is logical. This reasoning can be repeated at any position in the graph. For the relation between velocity and acceleration the same reasoning results in the red line which was drawn as  $-1$  times the sine function.

In fact by just looking at the graph, it is visually verified that the mathematics are correct. After all, when the position ( $x$ ) can be described as:

$$x(t) = \hat{x} \sin(\omega t) \quad (2.15)$$

then the speed ( $v$ ) equals the derivative of position over time:

$$v(t) = \frac{dx(t)}{dt} = \dot{x}(t) = \hat{x}\omega \cos(\omega t) \quad (2.16)$$

and the acceleration ( $a$ ) equals the derivative of velocity over time:

$$a(t) = \frac{dv(t)}{dt} = \frac{d^2x}{dt^2} = \ddot{x}(t) = -\hat{x}\omega^2 \sin(\omega t) \quad (2.17)$$

The visual representation is however not fully representative for the real situation, because it only gives the right answer on the amplitude of velocity and acceleration for a specific frequency when  $\omega = 2\pi f = 1$ . These amplitudes at another frequency can be found by straightforward reasoning, as a higher frequency means a faster change. For a constant position amplitude, the amplitude of the velocity increases proportional with the frequency and the amplitude of the acceleration with the frequency squared.

As a second step, the phase relationship between these three parameters is

important. The graph is just a time sample of an endless continuous signal and it could be concluded that velocity either advances on position with a phase of  $\phi = \pi/2$  [rad]  $\equiv 90^\circ$  or lags with a phase of  $\phi = 3\pi/2$  [rad]  $\equiv 270^\circ$ . The second option is however not logical as the position of an object is the result of **and thus comes after** the velocity. In mathematics this is noted in the following way:

$$x(t_1) = x(t_0) + \int_0^{t_1} v(t) dt \quad (2.18)$$

Where  $x(t_0)$  is the starting position at  $t_0$  and  $x(t_1)$  is the position at  $t_1$ . Likewise this is true for the acceleration that advances on the velocity. This means that the acceleration advances on the position with  $\phi = \pi$  [rad]  $\equiv 180^\circ$ .

### 2.1.3.2 Random signals or noise

The third kind of signals next to DC and AC is noise. Noise is a real random, non-deterministic signal which means that the signal value at any time of observation can have any value within a certain range. The range is described by means of statistical distributions that give the probability that the value is within that range. In Section 2.3 it will be explained that random signals in the time domain consist of an infinite amount of frequencies in the frequency domain. Also the distribution of these frequencies, called a *frequency spectrum*, can only be described in statistical terms. Noise is sometimes named with a colour term to indicate their frequency spectrum like *white noise* with an equal presence of all frequencies over the entire frequency spectrum or *pink noise* when the low frequency components are more stronger present than the high frequency components.

High performance Mechatronic systems exist by virtue of predictable and well controlled dynamics of the motion systems and to that respect random signals are by definition a part of the possible disturbances that impair the functionality. The only thing one can do is avoid or suppress the effects of noise but it can never be cancelled as that would require knowledge of future values and then the signal would no longer be random. The reduction of noise has always been an important area of research by determining the mechanisms behind the noise in order to improve the predictability and at least filter out the *systematic effects* that occur at regular intervals.

Noise and its statistical analysis is examined more in depth in Section 8.2 on *Dynamic Error Budgeting* where it will be explained how to deal with random disturbances in precision measurement and positioning systems.

### 2.1.3.3 Power of alternating signals

Sometimes electronic signals are represented by their power value, being the squared momentary value. This notation is for instance useful, when determining the impact on accuracy of different random disturbance signals on a system.

In case of a voltage signal, this power value would be equal to the power dissipated in a  $1 \Omega$  resistor, as in that case the current would be equal to the voltage. In an alternating signal the power also varies over time. With a purely resistive load, the power is always positive, with  $P_s \propto \sin^2(\omega t)$ . Although, as will be shown in several examples in this book, the power can be negative, when the current and the voltage have a different sign.

When examining the power of a signal, as defined by the squared momentary value, often the average value over time is taken as a representative number. This average value of the power  $P_s$  of a signal function  $f(t)$  is equal to:

$$\overline{P_s} = \frac{1}{T} \int_0^T \langle f(t) \rangle^2 dt \quad (2.19)$$

In the example of a voltage signal over a  $1 \Omega$  resistor this average power level would be equal to the power dissipated in the resistor by a DC voltage with a value of:

$$V_{\text{rms}} = \sqrt{\frac{1}{T} \int_0^T \langle f(t) \rangle^2 dt} \quad (2.20)$$

The term “rms” refers to Root Mean Square (RMS), named from the action of taking the root of the mean value of the squared function. The RMS value is a well-known term in electricity to characterise the useful value of the mains supply with a sinusoidal voltage  $V = \hat{V} \sin(\omega t)$ .

The equivalent DC voltage becomes:

$$\begin{aligned} V_{\text{rms}} &= \sqrt{\frac{1}{T} \int_0^T (\hat{V} \sin(\omega t))^2 dt} = \hat{V} \sqrt{\frac{1}{T} \int_0^T \sin^2(\omega t) dt} \\ &= \hat{V} \sqrt{\frac{1}{T} \int_0^T \frac{1 - \cos(2\omega t)}{2} dt} = \hat{V} \sqrt{\frac{1}{T} \left[ \frac{t}{2} - \frac{\sin(2\omega t)}{4\omega} \right]_0^T} \end{aligned} \quad (2.21)$$

By definition the RMS value has to be calculated over infinity to converge. Fortunately with repetitive periodic signals, the value is equal to the integration over  $n$  times the full period of the signal, where  $n$  is an integer. This

means that the sine term will average to zero and the final result becomes:

$$V_{\text{rms}} = \hat{V} \sqrt{\frac{1}{T} \left[ \frac{t}{2} \right]_0^T} = \hat{V} \sqrt{\frac{T}{2T}} = \frac{\hat{V}}{\sqrt{2}} \quad (2.22)$$

The 230 V AC mains supply in the Netherlands is the RMS value of the voltage so the peak value equals  $\hat{V} = \sqrt{2} \cdot 230 = 325$  V.

Similar to the voltage, the RMS value of a sinusoidal current will equal  $\hat{I}/\sqrt{2}$  and the combined power will equal  $V_{\text{RMS}} I_{\text{RMS}} = 0.5 \hat{V} \hat{I}$  as is proven with the following reasoning, when  $I(\omega) = \hat{I} \sin(\omega t)$  and  $V(\omega) = \hat{V} \sin(\omega t)$ :

$$P = I(\omega)V(\omega) = \hat{P} \sin(\omega t) \sin(\omega t) = \hat{P} \sin^2(\omega t) = \hat{P} \frac{1 - \cos(2\omega t)}{2} \quad (2.23)$$

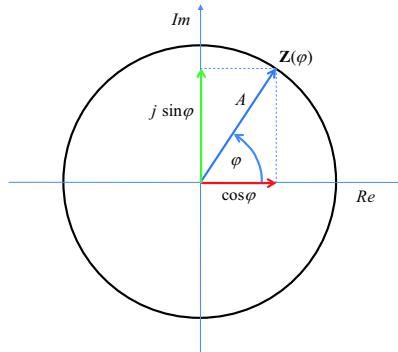
with the peak power  $\hat{P} = \hat{I}\hat{V}$ . The result is a positive number and averages out over time to  $P = \hat{P}/2$  because the average value of a sinusoidal function is zero.

#### 2.1.3.4 Representation in the complex plane

The dynamic behaviour of mechatronic systems is generally described in transfer functions for signals that act on the system. These functions are derived both in the time and frequency domain. The outcome of these transfer functions in the frequency domain describes the amplitude and phase of the output of the system in relation to the input as function of the frequency. It uses the mathematics of complex numbers because of its possibility to relate the phase to an angle in the complex plane as shown in Figure 2.7. The complex number  $Z = a + jb$  is represented in the complex plane as a vector with coordinates [a b]. In general not the vector notation but the following relation is used to describe a complex number with a phase angle  $\phi$ :

$$Z(\phi) = A(\cos \phi + j \sin \phi) = A e^{j\phi} \quad (2.24)$$

The exponential function is used in the mathematical derivation of several important transforms but for more understanding of the phenomena the sinusoidal and graphical representation is often used. As an example Figure 2.8 shows Ohm's law in a more generic representation where all values can be frequency dependant. For the load, this frequency dependency automatically leads to a phase shift between the load current and the voltage, as will be presented more in detail in the following chapters. For this reason the load is said to have an *impedance*  $Z$ , with often a complex



**Figure 2.7:** Vectorial representation of complex number  $Z(\phi) = A(\cos \phi + j \sin \phi) = Ae^{j\phi}$  where the phase shift is equivalent to the angle  $\phi$ .

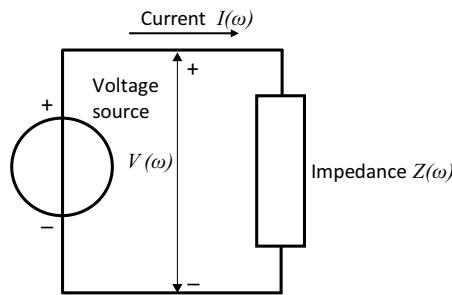
mathematical description like  $Z(\omega) = R + j\omega L$ , which makes it a load with a complex impedance.

In its generic form Ohm's law becomes as follows:

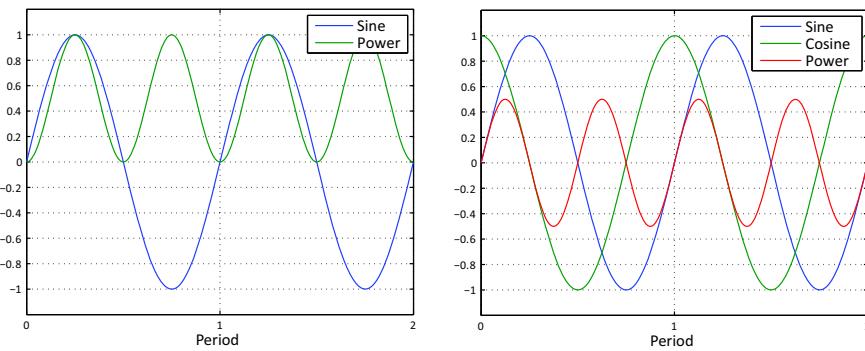
$$V(\omega) = I(\omega)Z(\omega) \quad \Rightarrow \quad I(\omega) = \frac{V(\omega)}{Z(\omega)} \quad (2.25)$$

Also the relation for power becomes less trivial as a phase shift between current and voltage has impact on the related power. Without a phase shift, the power was shown to average out over time to  $P = \hat{P}/2$ . In case of a  $\pi/2$  phase shift, where the current advances over the voltage, the Power becomes:

$$P = I(\omega)V(\omega) = \hat{P} \sin(\omega t - \pi/2)\sin(\omega t) = \hat{P} \cos(\omega t)\sin(\omega t) = \hat{P} \frac{1}{2} \sin(2\omega t) \quad (2.26)$$



**Figure 2.8:** Generic representation of Ohm's law. All variables can be frequency-dependent. The load becomes a “complex” impedance resulting in a possible phase shift between current and voltage.



**Figure 2.9:** Graphical representation of power as the multiplication of two sinusoidal functions. At the left side the functions are in phase, resulting in an average power of 0.5. At the right side the functions have  $90^\circ$  phase difference, resulting in an average power over time of zero.

This is a sinusoidal function with an average value of zero. The conclusion that can be drawn is that only the in-phase components of the current and voltage contribute to the average power into the load which is for instance the case with a resistor. The derived relation between power and phase is illustrated in Figure 2.9. When the phase difference is shifted gradually from zero to  $90^\circ$ , the average power shifts from 0.5 to zero.

It is important to note from the figure that, although the average power is zero, the momentary value of the power is not. At half of the time the power is positive, while in the other half the power is negative. This phenomenon of temporary energy storage occurs in many *reactive* impedances, as will be presented in the following chapters. In mechanics the mass of the body and the stiffness of a spring are reactive impedances, while the capacitor and inductor are the reactive impedances in an electronic system.

## 2.2 Energy propagation and waves

Directly related with frequency is the physical phenomenon of a wave. Waves are well-known from real life by the waves in water. In spite of this seemingly simple phenomenon, physical waves are not as straightforward to describe as it may seem. A valid first observation is, that waves deal with energy that is transferred to another place by a consecutive process that happens over the wave. This transfer is called *propagation* and happens with a certain velocity, which is called the propagation speed  $v_p$ . At a certain frequency, this propagation speed results in a spatial periodicity, that is called the *wavelength* with variable-symbol ( $\lambda$ ), related to the frequency by the following simple relation:

$$\lambda = v_p T = \frac{v_p}{f} \quad (2.27)$$

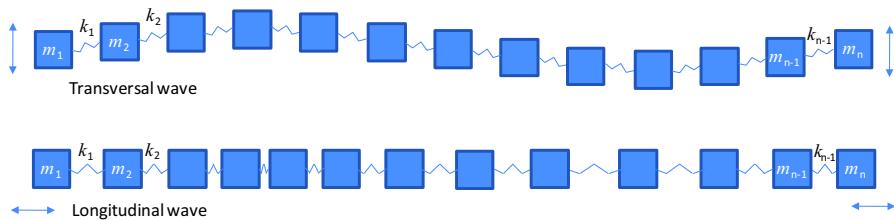
In the following, two essentially different types of waves will be presented, the real and observable mechanical waves, that are based on movements of material, and the physical-model-based electromagnetic waves. The examples are aimed to only create a basic idea of these phenomena that are sufficient to use it in this book. For that reason a full mathematical analysis with wave-equations is not presented.

### 2.2.1 Mechanical and acoustic waves

The propagation of **mechanical waves** through an elastic material can be explained with the help of a simplified ideal multi-body model consisting of a chain of springs and masses as shown in Figure 2.10. It is unsupported, floating in outer space and contains no damping parts. In the figure two different kind of waves are shown, **longitudinal** in the direction of the propagation direction and **transversal**, perpendicular to the propagation direction. In a three dimensional situation, the transversal movement can be in any direction orthogonal to the propagation direction. In the case of an electromagnetic wave, the transversal movement direction is called the *polarisation* direction.

To explain the principle of energy transfer, the longitudinal waves are taken as example, while the reasoning is also valid for transversal waves like in the example of the rope.

When a movement of mass  $m_1$  is introduced in the propagation direction of the chain, this will first cause a compression of the elastic coupling  $k_1$ .



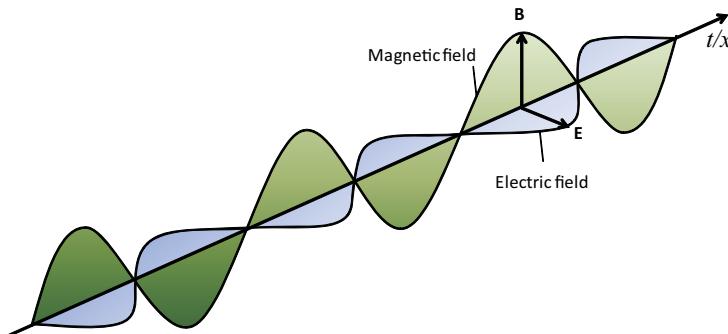
**Figure 2.10:** Transfer of energy by waves in a mechanical body can be either longitudinal or transversal. Transversal waves can be in any direction orthogonal to the propagation direction which is always longitudinal.

The resulting compression force is transferred to mass  $m_2$  which accelerates resulting in its own movement which causes in its turn a compression of  $k_2$ . This process is repeated over the total chain until the original movement reaches  $m_n$ . With this mechanism the kinetic energy from mass  $m_1$  is converted into potential energy in  $k_1$  which in its turn is transferred into kinetic energy of  $m_2$  and so on until the last mass is moving. This all under the assumption that no energy is lost. This phenomenon of transfer of energy in an elastic body is important in mechatronic systems because driving forces are also transported through the body as a wave and as a consequence will experience a delay between the actuator and the sensor when they are located separately. The propagation velocity  $v_p$  is determined by the elastic compressibility property, expressed by the Bulk elasticity  $B$ , and the density  $\rho$  of the medium and equals approximately:

$$v_p = \sqrt{\frac{B}{\rho}} \quad [\text{m/s}] \quad (2.28)$$

This equation is valid for gases and fluids and to a lesser extend for solid objects where it can be used only as approximation. The shape of the object also plays a role and the velocity for transversal and longitudinal waves is different. As an example with steel,  $B \approx 160 \cdot 10^9 \text{ N/m}$  and  $\rho \approx 8 \cdot 10^3 \text{ kg/m}^3$  resulting in a calculated velocity of approximately 4500 m/s. In reality the velocity in stainless steel varies between 3500 m/s for transversal waves and 5500 m/s for longitudinal waves. With for instance half a metre of steel this gives a delay of about one tenth of a millisecond which results in a phase delay of almost thirty-five degrees at one KiloHertz which can be significant from a control point of view!

**Acoustic waves** that relate to sound are a special subdivision of mechanical waves. In principle they can be visualised with the help of the same Figure 2.10. However with sound the medium is gas and this means that



**Figure 2.11:** Electromagnetic waves consist of a coupled periodic electric and magnetic field that propagate as function of time and position. In vacuum the velocity is equal to the speed of light  $c$ , exactly 299,792,458 m/s.

the mass elements are very small (molecules), while the elastic elements are also very small as they consist of the interaction between the moving molecules. On a macro-scale, the energy in acoustic waves is transferred by pressure and velocity where in practice only longitudinal acoustic waves exist because of the low viscosity, which could be the only mechanism to transfer transversal movements.

The example shown is still very much one dimensional and it is worthwhile to consider the three dimensional case at this point as the inserted energy is in reality transferred in all directions. To take this into account the transferred energy is defined as the energy per unit of surface. This value is often called *intensity*. Based on the law of conservation of energy this means that the intensity value will decrease as function of the enlarging surface which often in practice corresponds to something like a half sphere. When motion energy is inserted at the surface of an object, the intensity will decrease proportional with the squared distance from the point of insertion.

## 2.2.2 Electromagnetic waves

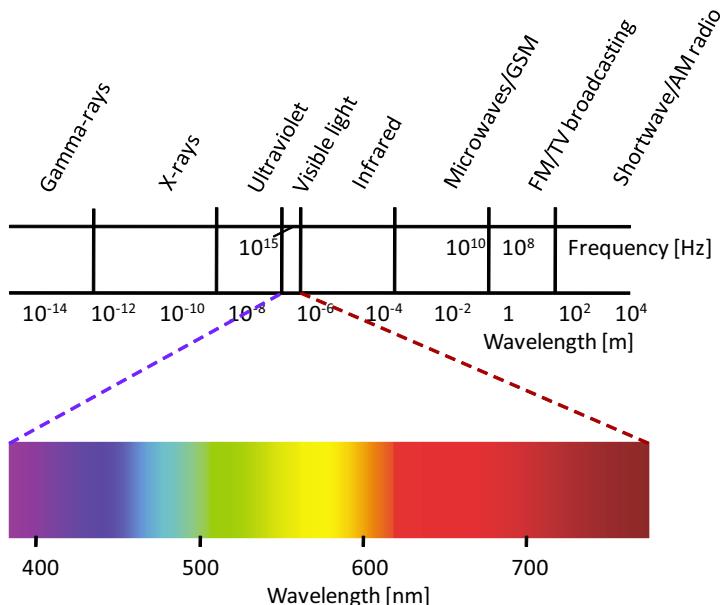
A special type of waves is the electromagnetic wave, because it does not use material for its propagation. Electromagnetic waves as a physical model were first postulated by James C. Maxwell (1831 – 1879) to describe phenomena like radio transmission and light. At first, people who were accustomed to the mechanical wave theory thought that electromagnetic waves should also have a carrier to propagate the energy. For this reason they invented an unknown medium called “Aether” as a wave carrier. Present theory

however states that electromagnetic energy is transferred by the interaction of electromagnetic fields. In fact this postulation of electromagnetic energy propagation as a wave is, like all physics, only a model to describe and predict behaviour of a physical phenomenon without explaining anything. It is only valid under certain constraints. The constraints of the electromagnetic wave theory only apply when the energy levels are so small that the energy transfer starts to behave as separate single units of energy (quanta) that were first postulated by Max K.E.L. Planck (1858 – 1947). The relating theory of Quantum Electrodynamics on the interaction of photons and electrons has been described by Richard P. Feynman (1918 – 1988) in his well readable book “QED, The strange theory of light and matter”. It gives a calculation model for electromagnetic phenomena by means of probability statistics. Though it is in itself very interesting, especially to become aware of the futility to search for answers **why** nature does the weird things it does, reading it is not necessary to be able to work with electromagnetic phenomena at the scale used in mechatronic systems. The wave theory is mostly sufficient for this purpose. Nevertheless it will be shortly referenced in the optics chapter, when presenting diffraction on gratings.

Figure 2.11 shows the transversal wave model of an electromagnetic wave as a combination of two orthogonal periodic fields, a magnetic (**B**) and an electric (**E**) field. This model is used in Chapter 7 on optics to describe phenomena like interference and diffraction. Electromagnetic waves cover a very wide spectrum as shown in Figure 2.12. It starts with the Gamma radiation or Gamma rays at one extreme of the spectrum and radio waves at the other. The important area of visible light is found around a wavelength of 600 nm. Electromagnetic waves propagate with the speed of light  $c$ , which is in vacuum exactly 299,792,458 m/s. This enormous velocity results in an extreme unmeasurable temporal frequency of  $\approx 5 \cdot 10^{14}$  Hz for visible light.

### 2.2.2.1 Transferred energy and amplitude

It has been mentioned that waves represent the transfer of energy by means of an oscillating medium. In the mechanical wave this energy is alternatingly present in the form of kinetic energy or as potential energy. When looking at the maximum kinetic energy  $E_k = 1/2mv^2$  at any point in a mechanical wave it is clear that the related intensity is proportional to the amplitude squared. Also with electromagnetic waves the intensity is proportional to the amplitude squared, but with these waves generally the term *irradiance* is used rather than intensity because in radiology the term intensity or *radiant intensity* is also used as the radiated power per unit of solid angle

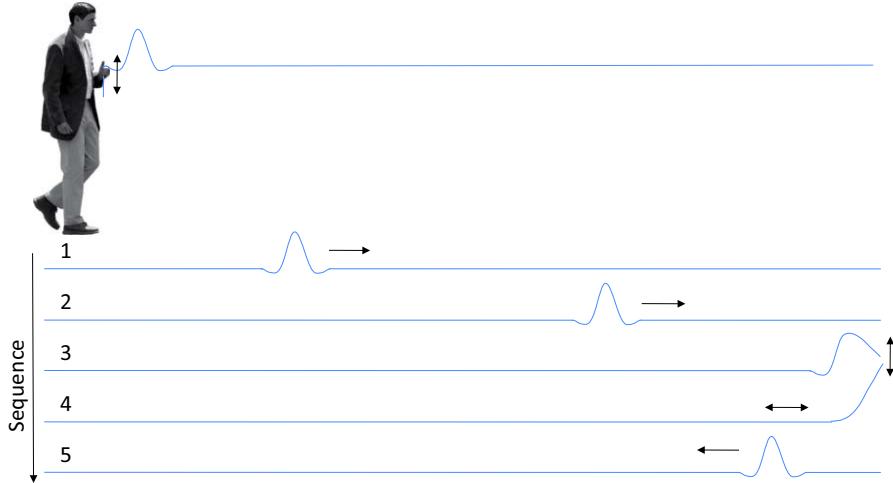


**Figure 2.12:** Electromagnetic waves are observed over a vast spectrum of frequencies, ranging from radio waves to gamma rays. The spectrum of visible light covers only a very small part.

originating at the source of the wave. It is typical for waves that start at a singular spot. The term irradiance will be frequently used in the chapters on optics and measurement.

### 2.2.3 Reflection of waves

A rope is an example of a mechanical chain consisting of an infinite number of interconnected masses and springs. Figure 2.13 shows a situation where a man inserts a sequence of an upward and equal downward movement in a rope causing a transversal half wave. The rope is not connected to anything at its end point and the effects of gravity are neglected. When the wave arrives at the end point, the momentum can not be transferred further to a next neighbouring part of the rope. This zero force boundary condition at the end point results in a continuation of its upward movement. This lasts until the originally driving force from the preceding part of the rope starts to slow down the upward movement of the end point, while transferring the momentum back to the previous part of the rope. This effect is called *reflection* and the resulting backwards moving half wave has the same sign

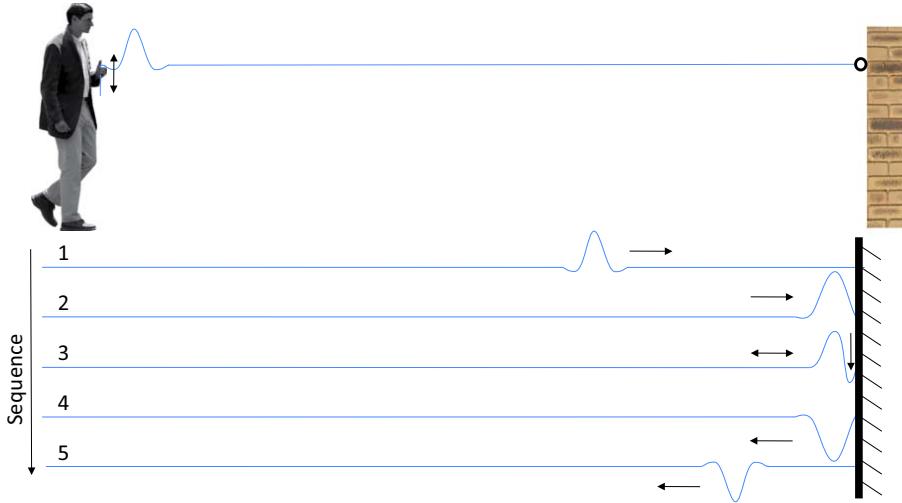


**Figure 2.13:** A man holds a rope with a loose end and inserts a half wave by an upwards movement followed an equal downward movement. The reflection of the wave at the loose end has the same sign because of the zero force boundary condition at the end point. The effect of gravity is neglected.

as the original half wave, because it first starts with an upward momentum, originating from the extreme upward motion of the end, followed by a downward movement.

Another situation occurs when the end of the rope is connected to a rigid stationary object like a wall. This can in reality be visualised also with presence of gravity and is shown in Figure 2.14. This zero movement boundary condition causes the force at the connection point to increase until all energy in the wave is absorbed in the local elasticity in the rope. That potential energy is inserted back in the rope such that its movement gains a maximum momentum passing the centre line until its slowed down again at the downward side where its energy will be transferred in the backward direction. The reflection in this situation leads to a sign reversal, because it first starts with an downward momentum, originating from the extreme force at the connection point, followed by an upward movement.

Of course these two examples can also be calculated with real mathematics. This would enable to precisely calculate the transfer over a discontinuity and show the minus sign in the reflection at a rigid connection but as mentioned, is not necessary for a sufficient comprehension of the phenomena in the

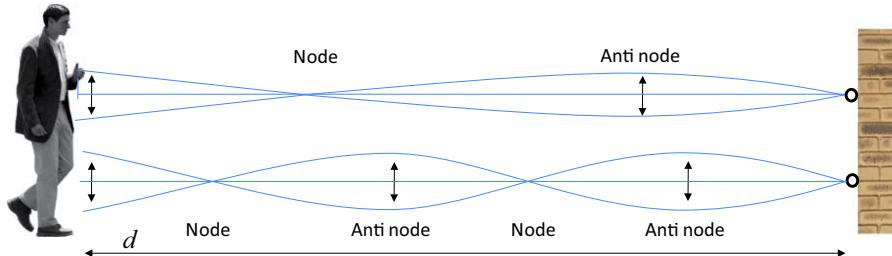


**Figure 2.14:** The reflection of a movement at a rigidly connected end results in a sign inversion of the reflected wave due to the zero movement boundary condition. In this situation the force at the connection point will not only stop the transversal wave but reflect it in the downward direction.

scope of this book. This phase reversal is also observed with the reflection of electromagnetic waves when going from a low density to a high density medium like with reflecting on a mirror.

### 2.2.3.1 Standing waves

A special situation occurs when the man inserts a sinusoidal movement in the rope with such a frequency that the reflected wave arrives in phase with the exerted wave. This situation is shown in Figure 2.15. The connection point at the wall leads to sign reversal and the side of the man to the same phase. Sign reversal of a sinusoidal wave is equal to a  $180^\circ$  or  $\lambda/2$  phase shift, which means that at least another  $180^\circ$  phase shift is necessary to be in phase again with the inserted movement. This is accomplished when the wave travels a distance of at least  $\lambda/2$ . because the wave travels twice the distance from the man to the wall the minimal distance becomes  $\lambda/4$ . But also with a distance of  $d = 3 \cdot \lambda/4$  an in phase reflection is obtained, because a path of two times  $2 \cdot \lambda/4$  equals a full wavelength, giving a  $360^\circ$  phase shift. This can be written in the following generic way, when using the relation



**Figure 2.15:** A standing wave occurs with a sinusoidal excitation when the reflected wave is in phase with the excited wave. This occurs at a frequency equal to  $f = nv_p/4d$  with  $n = 1, 3, 5, 7, \dots$ . The situation for  $n = 3$  is shown in the upper graph and  $n = 5$  for the lower graph.

between frequency and wavelength of Equation (2.27):

$$f = \frac{v_p}{\lambda}, \quad d = \frac{n\lambda}{4} \quad \Rightarrow \quad f = \frac{nv_p}{4d}, \quad n = 1, 3, 5, 7, \dots \quad (2.29)$$

Due to the in phase reflection at these frequencies the velocity at the point of insertion will also be in phase with the force by the man, achieving a continuous energy transfer to the rope and consequently the motion amplitude will increase. This effect is called a *resonance* as will be also presented in Chapter 3 on dynamics. In the figure, fixed positions are shown, where the amplitude is zero and those positions are called a “node”. At these positions the energy is stored in elastic deformation only. The locations where the motion amplitude is maximum are called the “anti-node” and the energy is stored in kinetic energy at the highest velocity moment when the rope passes the centre line.

It is important to notice that the excitation force has to be inserted at an anti-node to be effective. This is true because insertion at a node with zero velocity will not transfer energy and at an anti-node the velocity will be maximal. This effect is used in any musical snare-instrument where the musician exerts forces in an anti-node on the snare to create the sound. Depending on the position of exertion a different tonal balance is achieved because with each frequency another place will correspond with an anti-node. In a guitar the sound contains more high frequencies when played closer to the bridge, the connection of the snare with the instrument. Close to the bridge the anti-nodes of only these high frequency harmonics are located which is also observed in Figure 2.15 where the wall corresponds with the bridge.

## 2.3 Mathematical analysis of signals and dynamics

Several mathematical methods are used to model dynamic phenomena and signals in mechatronic systems. In this section two related mathematical transform principles are presented, the Fourier and the Laplace transform that are frequently used for this purpose

### 2.3.1 Fourier transform

In real mechatronic systems a sinusoidal movement of one single frequency can only occur at a resonance frequency. Real life is full of periodic events with many different frequencies that occur simultaneously without any correlation. Fortunately the French mathematician Jean Baptiste Joseph Fourier (1768 – 1830) determined that any periodic signal can be seen as a combination of sinusoidal signals with a clear harmonic frequency interrelation. This is true under the condition that the function contains no discontinuities as these would lead to a non-converging series. Due to the natural source of the periodic events of interest, these conditions are always met.

The mathematical method is called the *Fourier transform* because it transforms a function in the time domain to a function in the frequency domain. It consists of a series expansion and is based on the following trigonometric identity:

$$\sin(\omega_1 t) \sin(\omega_2 t) = \frac{-\cos(\omega_1 t + \omega_2 t) + \cos(\omega_1 t - \omega_2 t)}{2} \quad (2.30)$$

When  $\omega_1 = \omega_2$  the first term in the numerator becomes a standard cosine function averaging to zero over time and the second term in the numerator becomes equal to one resulting in a non-zero (0.5) total average of the function. In case  $\omega_1$  and  $\omega_2$  are not equal both terms in the numerator would become simple cosine functions both averaging to zero over time. Based on this sharp difference between the two situations, the average value over time of the multiplication of an arbitrary periodic function with a sinusoidal function of a certain frequency would be zero, unless the arbitrary periodic function would contain a frequency component equal to, and in phase with, the sinusoidal function.

With this mathematical conclusion the Fourier series of a periodic function is calculated as follows. First the formula will be applied in the form where

the value is a function of the angle  $\theta$  instead of time and the interval of one period  $T$  is represented as an angle from  $-\pi$  to  $\pi$ . When  $F(\theta)$  represents the Fourier transform of  $f(\theta)$  then:

$$F(\theta) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\theta) + b_n \sin(n\theta)] \quad (2.31)$$

Where  $n$  is an integer  $n \geq 1$ . Because the angle  $\theta$  equals  $\omega_0 t$  where  $\omega_0$  represents the fundamental frequency of the periodic function in the time domain, the Fourier transform of this temporal periodic function can also be written as:

$$F(t) = a_0 + \sum_{n=1}^{\infty} [a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)] \quad (2.32)$$

The sine and cosine terms define the phase relation between the different components. The Fourier coefficients  $a_n$  and  $b_n$  are the amplitudes of the corresponding sine and cosine terms belonging to the different frequencies and are calculated over the mentioned interval by the next relations:

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) d\theta \quad (2.33)$$

This first coefficient represents the average value and is zero for a periodic signal without a DC offset. The other terms become according to the above mentioned trigonometric identity:

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \cos(n\theta) d\theta \quad (2.34)$$

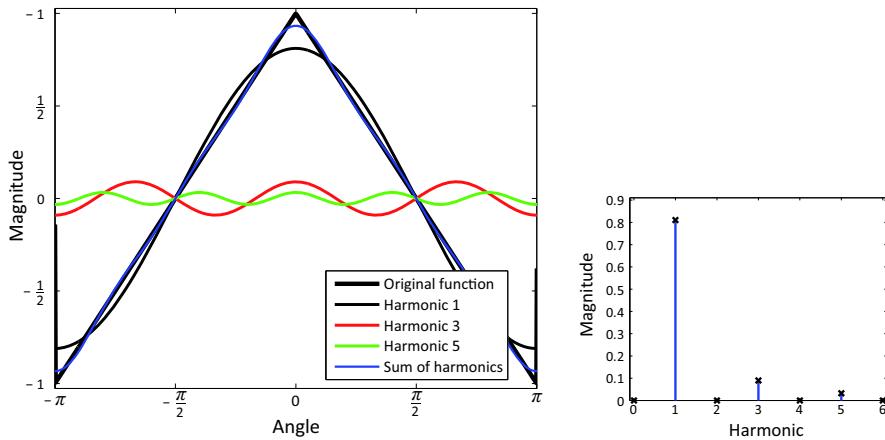
and:

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \sin(n\theta) d\theta \quad (2.35)$$

In equation (2.32) the frequency terms are all an integer multitude of the fundamental frequency. This interrelation is called harmonic because in music a waveform that consists of only harmonically related frequencies does sound like a harmonic (not out of tune) tone. The fundamental frequency equals the first harmonic and likewise the frequencies corresponding with  $n = (1, 3, 5, \dots)$  are called odd harmonics and the frequencies corresponding with  $n = (2, 4, 6, \dots)$  are called even harmonics<sup>2</sup>. To illustrate and underline

---

<sup>2</sup>The terms “odd” and “even” should not be confused with odd and even functions in mathematics that deal with symmetry relations of functions. A function is called even when the graphic representation in the  $x - y$  plane is unchanged when mirrored around the  $y$  axis. The graphic representation of odd functions remain unchanged when rotated  $180^\circ$  around the origin. For example  $\cos x$  is an even function while  $\sin x$  is an odd function.



**Figure 2.16:** Triangle waveform, approximated by the sum of the first three harmonic terms of the corresponding Fourier series, already comes close to the ideal waveform.

this theory on Fourier series of periodic functions the following three waveforms are taken as example, the triangle, sawtooth and square waveform.

### 2.3.1.1 Triangle waveform

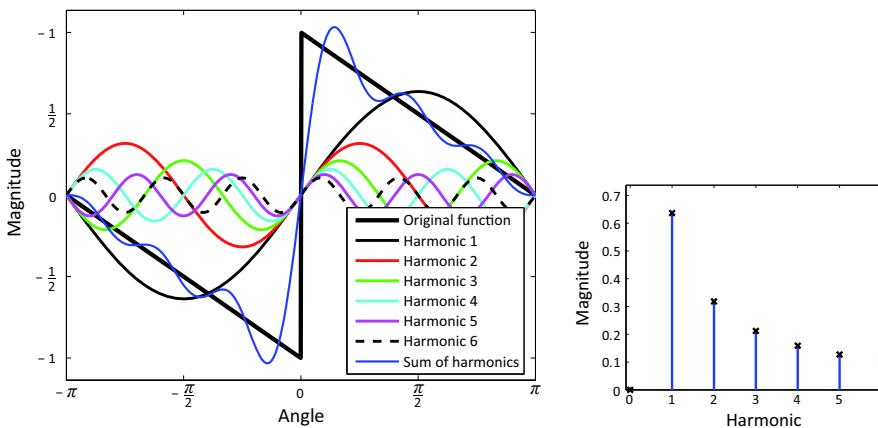
The triangle waveform is shown in Figure 2.16. It is clear that  $a_0$  is zero as the waveform has an average value over time of zero. Also the fundamental frequency is most probably a cosine function as it is like the cosine an even function. The Fourier terms can be calculated from the mathematical expression that describes the triangle waveform as function of the angle  $\theta$ .

$$\begin{aligned} f(\theta) &= 1 + \frac{2}{\pi}\theta && \text{for } -\pi < \theta \leq 0 \\ f(\theta) &= 1 - \frac{2}{\pi}\theta && \text{for } 0 < \theta < \pi \end{aligned} \quad (2.36)$$

When calculating the coefficients of the Fourier series of this function in time notation ( $\theta = \omega t$ ), the following infinite series of frequencies is found:

$$F(t) = \frac{8}{\pi^2} \left( \cos(\omega t) + \frac{1}{3^2} \cos(3\omega t) + \dots + \frac{1}{n^2} \cos(n\omega t) \right) \quad n = 1, 3, 5, 7, \dots \quad (2.37)$$

In the figure indeed the three frequencies all have a maximum at  $\theta = 0$ , which means that by summing them up, the top pushes up to shape a



**Figure 2.17:** Sawtooth waveform, approximated by the sum of the first six harmonic terms of the corresponding Fourier series, is still hardly recognisable as a real sawtooth.

sharp tip if all terms would be included. It is also nice to see that between for instance  $\theta = \pi/4$  and  $\pi/2$  the combination of the second and third term leads to a flattening of the curved slope of the first term in that region. The mathematics also show, that only odd harmonics are present. Due to the quadratic term in the numerator of the amplitude coefficients the higher order terms are rapidly becoming negligible. The image in the figure consists of only three terms from the Fourier series and the sum already looks like a triangle. Nevertheless due to the sharp edges an infinite amount of harmonics would be needed to create an ideal triangle waveform.

### 2.3.1.2 Sawtooth waveform

The sawtooth waveform is shown in Figure 2.17. While it might look like another kind of triangle it is not that simple. A theoretical ideal sawtooth waveform has an infinitely steep slope around  $\theta = 0$  which represents a discontinuity. As a consequence many terms are required to approximate a sawtooth waveform to an acceptable level of accuracy. Like with the triangle  $a_0$  is zero but in this case the fundamental frequency of the range is a sine function as the sawtooth is like the sine an odd function. The mathematical relation that describes the sawtooth waveform as function of the angle  $\theta$  is

as follows.

$$\begin{aligned} f(\theta) &= -1 + \frac{1}{\pi}\theta && \text{for } -\pi < \theta < 0 \\ f(\theta) &= 1 - \frac{1}{\pi}\theta && \text{for } 0 < \theta < \pi \\ f(\theta) &= 0 && \text{for } \theta = 0 \end{aligned} \quad (2.38)$$

Calculating the coefficients of the Fourier series of this function in time notation ( $\theta = \omega t$ ), gives the following infinite series of frequencies:

$$F(t) = \frac{2}{\pi} \left( \sin(\omega t) + \frac{1}{2} \sin(2\omega t) + \dots + \frac{1}{n} \sin(n\omega t) \right) \quad n = 1, 2, 3, 4, \dots \quad (2.39)$$

This means that both odd and even harmonics are present. Furthermore, the amplitude coefficients of the successive harmonics only decrease proportional with  $n$  which underlines that many terms are required to create a reliable sawtooth. This is due to the above mentioned fact that the infinitely steep ramp has to be created by summation of only finitely steep sine wave slopes. The image in the figure consists of only six terms and though one faintly recognises the final shape it is still far from a sawtooth waveform.

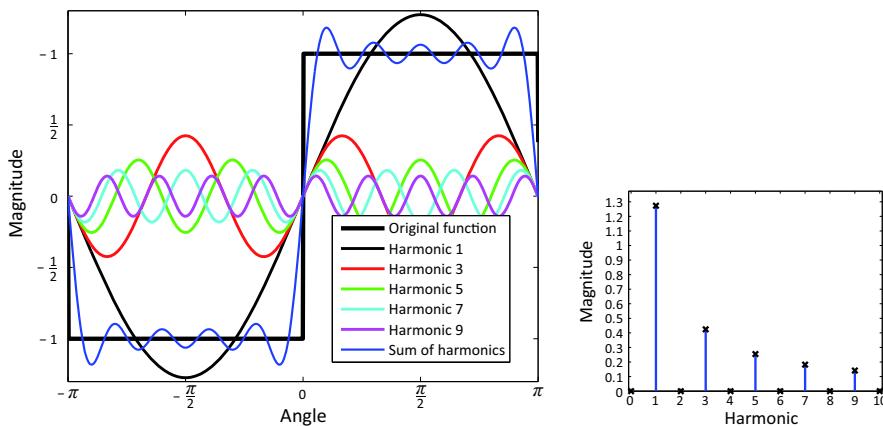
### 2.3.1.3 Square waveform

The third example is the square waveform as shown in Figure 2.18. In this case two discontinuities are present in the signal, which also means that more terms are required to approximate the waveform with sinusoidal functions. Like with both previous waveforms  $a_0$  is zero and like with the sawtooth also the fundamental frequency of the range is a sine function as it is an odd function. The mathematical relation that describes the square wave as function of the angle  $\theta$  is as follows.

$$\begin{aligned} f(\theta) &= -1 && \text{for } -\pi < \theta < 0 \\ f(\theta) &= 1 && \text{for } 0 < \theta < \pi \\ f(\theta) &= 0 && \text{for } \theta = -\pi, 0, \pi \end{aligned} \quad (2.40)$$

Calculating the coefficients the Fourier series of this function in time notation ( $\theta = \omega t$ ) gives the following infinite series of frequencies:

$$F(t) = \frac{4}{\pi} \left( \sin(\omega t) + \frac{1}{3} \sin(3\omega t) + \dots + \frac{1}{n} \sin(n\omega t) \right) \quad n = 1, 3, 5, 7, \dots \quad (2.41)$$



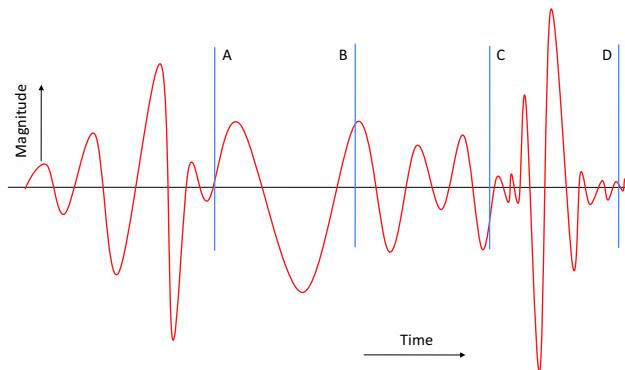
**Figure 2.18:** Square waveform, approximated by the sum of the first five harmonic terms of the corresponding Fourier series starts to look like a real square.

This means that like with the triangle waveform only odd harmonics are present. Furthermore, like with the sawtooth waveform, the amplitude coefficients of the successive harmonics only decrease proportional with  $n$ , which underlines that also in this case many terms are required to realise a reliable square wave for the same reason as an ideal theoretical square waveform has discontinuities that can only be approximated. The fact that the summed first six terms of the Fourier series already looks more like a real square waveform than the previous example is due to the missing even harmonics leading to a higher frequency of the last terms that were included in this approximation.

Based on these three examples the following observation can be made. A periodic signal that is symmetrical around an axis in the  $y$  direction at  $\pi/2 \pm n\pi$  will create only odd harmonics. The even harmonics all represent an asymmetry around these axes as can be seen with the sawtooth waveform. This is especially important with spatial frequencies in optics where in principle only odd harmonics are present.

#### 2.3.1.4 Fourier analysis of non-periodic signals

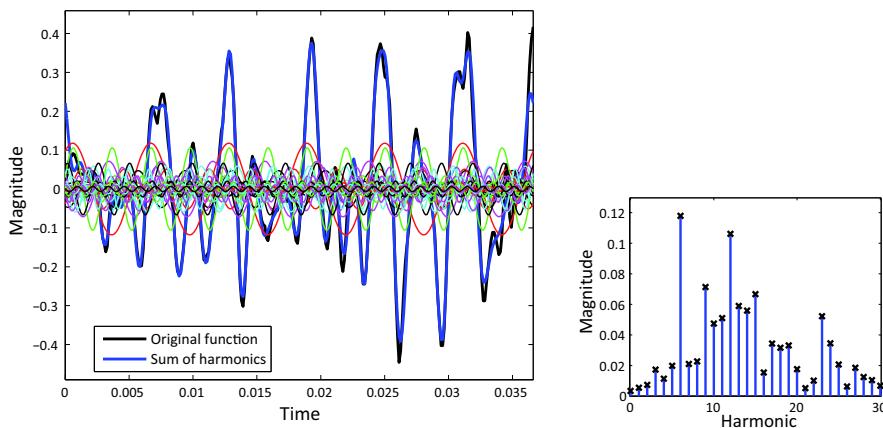
From the previous part it was demonstrated, that a periodic signal can be decomposed into harmonically related frequencies. The process of calculating the terms in the equation focused on one period only as being representative



**Figure 2.19:** Graphical representation of a random waveform in the time domain.

for the entire continuous function. In reality signals often look more like the graph of Figure 2.19 showing a random signal that was registered over a short period of time. A real random signal consists of an infinite amount of different sinusoidal frequencies with different (and changing!) amplitudes and phase. When only a limited time span is observed, like shown in the figure one can recognise areas with a low frequency like in area A – B and with a high frequency like the area C – D. It is possible to approximate such a random signal by a limited amount of frequencies when observing only a limited time span. Figure 2.20 shows an example that is created by a further execution of the above mentioned Fourier transform methodology and it works as follows:

A sample of the signal is taken over the defined time span and this sample is multiplied with a series of sine and cosine frequencies covering the spectrum of interest. For example in mechatronic systems a spectrum of 0 – 1000 Hz could be taken with steps of 1 Hz resulting in two thousand multiplications. The outcome of each multiplication results in the amplitude of each of these frequencies and can be represented graphically in the form of a frequency spectrum. It is clear that a lot of computation power is necessary to do this with a large resolution in frequencies especially when real-time information of fast changing signals with many samples is needed. Even with the fast computing systems of today this *discrete Fourier transform* (DFT) is not used anymore as the same result can be achieved by means of a more efficient algorithm called *Fast Fourier Transform* (FFT). This method of which the mathematical details go beyond the scope of this book has become the standard for signal analysis and is for instance applied in instruments like a *Spectrum Analyser*. Next to the digital FFT operation, Spectrum



**Figure 2.20:** A random waveform approximated by the sum of several harmonic terms of the corresponding Fourier series. It shows that even a random signal observed over a limited time span can be reliably represented by a combination of sinusoidal waveforms.

Analysers can also be based on analogue mixing of the unknown signal with a known frequency and creating a multiplication by non-linearity in the amplification. This method is used in very high frequency analysers.

### 2.3.2 Laplace transform

The Laplace transform is a mathematical method that was created by the French astronomer and mathematician Pierre-Simon, Marquis de Laplace (1749 – 1827). This transform is very useful for the analytical description of dynamic systems as it enables to solve differential equations in the *time domain* by solutions in the *frequency domain*. It is very much related to the Fourier transform and will be used throughout several chapters in this book, for which reason it is included in this chapter.

In case  $f(t)$  is a function of the time variable  $t$ , then the Laplace transform  $f(s) = \mathcal{L}\{f(t)\}$  is described as a function of the Laplace variable  $s$  in the following way:

$$f(s) = \mathcal{L}\{f(t)\} = \int_0^{\infty} e^{-st} f(t) dt \quad (2.42)$$

The Laplace variable  $s$  is a complex number:

$$s = \sigma + j\omega \quad (2.43)$$

where  $j^2 = -1$  (also often noted as  $i$ ) and  $\sigma$  and  $\omega$  are real numbers. A complete treatise on all aspects of this transform falls beyond the scope of this book. The most important result of this transform for time varying functions is the possibility to replace the differential over time by the Laplace variable  $s$  and by  $j\omega$ . When investigating the frequency response of a dynamic system, the real number  $\sigma$  can be neglected but it will be used in examining the *poles and zeros* of a transfer function, as will be presented in Chapter 3 on dynamics and Chapter 4 on motion control.

The Laplace transform converts a differential of a variable  $x(t)$  over time into the following expression in the frequency domain:

$$\mathcal{L} \left\{ \frac{dx(t)}{dt} \right\} = sx(s) = j\omega x(s) \quad (2.44)$$

The same transform can be applied to an integration action which gives the following result:

$$\mathcal{L} \left\{ \int_0^t x(t) dt \right\} = \frac{x(s)}{s} = \frac{x(s)}{j\omega} = -j \frac{x(s)}{\omega} \quad (2.45)$$

it can be concluded that the  $s$  term due to the differentiation over time of a variable results both in a proportional increase of the magnitude of the variable with increasing frequency by the multiplication with  $\omega$  and in a positive phase shift of  $90^\circ$  corresponding with the positive imaginary term  $j$ . Following the same reasoning, the integration of a variable over time gives both a proportional decrease of the magnitude of the variable with increasing frequency because of the division with  $\omega$  and a phase shift of  $-90^\circ$  corresponding with the negative imaginary term  $-j$ .

The frequency domain, when written with the  $s$  variable, is also frequently called the *Laplace domain*. It is important that equations from the frequency domain are not confused with equations from the time domain. When it is necessary to clearly distinguish the different domains, the notation of a function like  $f(t)$ , with the variable  $t$  between round brackets, strictly refers to the time domain. The notations with the variable  $(s)$  for the Laplace domain,  $(\omega)$  for the radial frequency or  $(f)$  for the temporal frequency all strictly refer to the frequency domain. The addition of the distinguishing variable terms  $(t)$ ,  $(s)$ ,  $(\omega)$  or  $(f)$  will in many cases be only done once in every equation at the term(s) before the equal sign in order to avoid too many

brackets in the equations. This limited notation is allowed and sufficient as an equation is uniquely valid, either when the entire equation is defined in the frequency domain or when the entire equation is defined in the time domain.

## 2.4 Dynamic system response to a stimulus

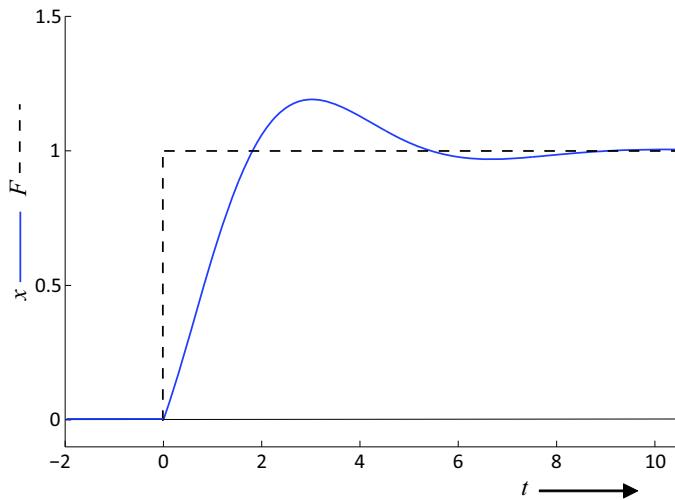
In the previous part of this chapter signals are introduced in terms of frequencies. As a next step it is necessary to link this knowledge to the analysis of the dynamic behaviour of mechatronic systems both in the time and frequency domain. This implies the introduction of the term *stimulus* as being a specific signal, created with the purpose of stimulating a reaction from a dynamic system. In most cases it is a force created by actuators or by external vibrations.

The experimental investigation into dynamic properties of a mechanical system is always done by applying external forces on different locations of the system and by simultaneous measurement of the response of the system to these stimuli. The frequency content of these stimuli can be quite different. Depending on their origin, several typical responses to standardised stimuli are distinguished of which the most important are:

- **Step Response** to a unit step stimulus.
- **Impulse Response** to a unit impulse stimulus.
- **Frequency Response** to a stimulus with a wide frequency spectrum that is continuously available.

### 2.4.0.1 Step response

The step response is a typical time domain related phenomenon and generally the most straightforward of all responses. It is often quite easy to understand or to visualise what happens when a force on an object increases from zero to a certain value. For instance when putting a heavy load in a car one sees the car sagging with a certain speed and depending on the suspension and the shock absorbers (dampers), a slow periodic movement with decreasing amplitude is observed. In Figure 2.21 a typical response of a dynamic system to a sudden change of force (step) from value 0 to 1

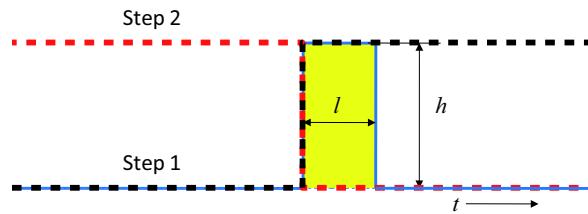


**Figure 2.21:** Typical response of a damped mass spring system to a force on the mass that changes from 0 to 1 at  $t = 0$ . The vertical axis is normalised to 1 at  $t = \infty$ .

is shown. The dynamic system in this case is a damped mass spring system like presented in Chapter 3. This response, if considered upside down, corresponds to the described response of the car.

Even with the origin of the step response in the time domain its equivalent in the frequency domain can be derived by the Fourier transform. This shows that a step function consists of a continuous spectrum of sinusoidal frequencies starting at 0 Hz that all are in phase at the moment of the step and which have an amplitude that is inversely proportional to the frequency. This can be imagined when one considers that all those frequencies will only be correlated (in phase) at  $t = 0$  but at any other time their correlation is lost and the combination will average out to zero. In a mathematical sense a step function is a special case of a square waveform with a fundamental frequency of 0 Hz. Starting with the Fourier series of a square waveform according to Equation (2.41), the limit of  $\omega \rightarrow 0$  results in a continuous spectrum of frequencies starting at 0 Hz.

Because of this wide frequency spectrum, a step function is a very suitable stimulus **for a passive mechanical system** to investigate the occurrence of resonances. It is important to note however that the step response is less applicable as stimulus in an actively controlled dynamic system with actuators as with most controllers the infinite slope will force the actuator into saturation and cause other non-linear effects that are not observed



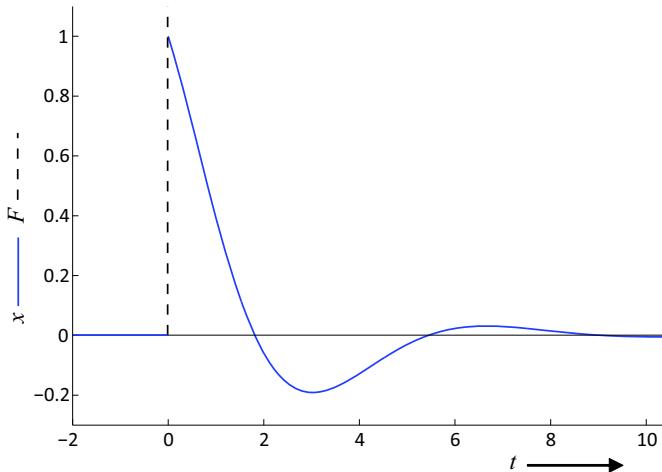
**Figure 2.22:** An impulse function as a combination of two steps. In case of a force step the surface defined by  $l$  and  $h$  equals 1 Ns.

under normal conditions.

Although a step is a one time event, one can as alternative approximate the step by a low frequency square wave. This limits of course the spectrum to the corresponding harmonic frequencies but depending on the frequency range of interest this can help in doing multiple measurements to reduce the impact of random noise which will be averaged out.

#### 2.4.0.2 Impulse response

An **impulse function** is a special combination of two steps, one up and one down directly after each other as shown in Figure 2.22. The ideal impulse lasts for an infinitely short time with an infinitely high magnitude with a multiplied value of one. The response on such a stimulus of the same dynamic system as was used with the step response is shown in Figure 2.23. This impulse response looks quite similar to the corresponding step response although it starts at  $x = 0$  and stabilises on  $x = 0$  again. The main reason why this stimulus is so important and widely used is the fact that in the frequency domain it contains all frequencies *at equal amplitude*. Measuring the response to an impulse and performing a FFT analysis of this response immediately shows all resonances of a dynamic system. Unfortunately it is not possible to create an ideal impulse function in reality, due to limitations at high frequencies of the physical embodiment.



**Figure 2.23:** Typical response of a damped mass spring system on a force impulse at  $t = 0$ . The vertical axis is normalised to 1 just after the impulse.

This means that the length becomes longer than zero and the peak value proportionally lower than infinite while simultaneously the edges are rounded off. To a certain extent one can compensate the corresponding lack of high frequencies by convolution<sup>3</sup> of the measured response with the real spectrum of the stimulus but at higher frequencies the noise will limit the reliability of the outcome.

The impulse response is often used for modal analysis of a mechanical system. The impulse is applied by means of a standardised "hammer". The hammer is provided with a rubber tip to avoid damage by the "infinite" force and as a consequence, the impulse will be rounded off with a lower amplitude of the high frequencies. By measuring the exerted impulse itself directly on the hammer, the deviation of the ideal pulse can be taken into account.

One important drawback of this stimulus is that all the energy is inserted in the system in a very short time with a high force over a short time. Inevitably this will give erroneous information in case of non-linear systems. As with the step response, for this reason also the impulse is less suitable for active controlled dynamic systems.

---

<sup>3</sup>Convolution is a mathematical method to define the cross-correlation between different functions.

### 2.4.0.3 Frequency response

Frequency domain related stimuli avoid the extremely high forces that are related to the “single event stimuli” of the previous section, because they present a more or less permanent continuous spectrum of frequencies with sufficiently reduced amplitudes. These signals enable longer measurements without errors due to nonlinearity.

Sinusoidal waveforms are quite easy to create in electronics. For that reason, initially stimuli were used that consisted of a sinusoidal waveform with a constant amplitude and a gradually changing frequency. This “frequency sweep” was suitable for fast electronic systems. In mechanical systems with inherent lower frequency resonances the frequency sweep method sometimes induced errors due to the time needed for resonances to gain sufficient energy to manifest themselves. A fast sweep would cause these resonances to remain unnoticed. One way to overcome this problem is using a very slow sweep frequency if time allows and this process could be made interactive to zoom in on critical areas. A better method is to use a stimulus containing all frequencies over a longer time than with an impulse. The first example is to use *white noise* as stimulus of which the power (signal value squared) is constant for any fixed frequency range over the total frequency band. In practical situations, similar to the impulse, ideal white noise does not exist due to high frequency limitations in the physical embodiment and only the frequency band of interest is excited.

The second example of a more continuous frequency spectrum is created in this era of high speed digital electronics and is based on the synthesis of a stimulus signal by means of *multi-sines*. In principle this approach allows to create any spectrum of signals, with correlated phase and frequency, random phase relations or any other combination. Frequency areas of no interest can be avoided, different amplitudes can be chosen for different frequency areas to avoid overheating or non-linear behaviour and the frequency spectrum can be refined in the areas of most interest. As the signals are fully deterministic, all phase and amplitude relations in the stimulus are known and it becomes fully straightforward to determine the dynamic system response.

## 2.4.1 Graphical representation in the frequency domain

Two representations are frequently used to display the response of a dynamic system to these frequency domain related stimuli:

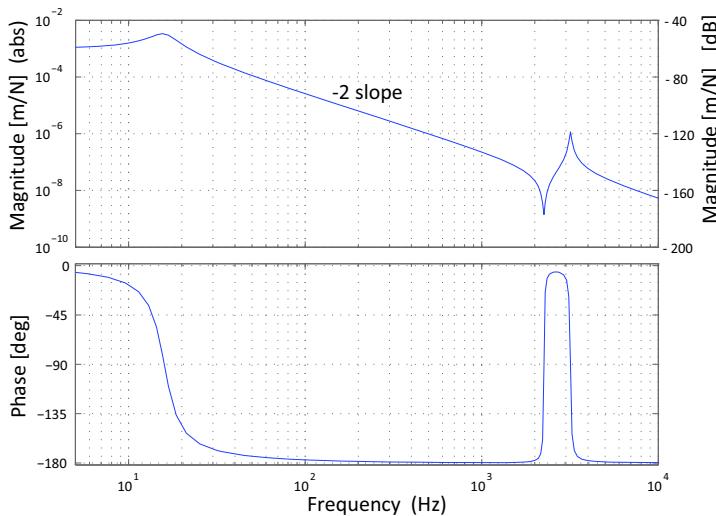
- The **Bode-plot** to show the magnitude and phase response upon a continuous frequency stimulus in two graphs as function of the frequency, one for the magnitude and one for the phase.
- The **Nyquist plot** to show the magnitude and phase response on a continuous frequency stimulus in one graph as function of the frequency.

### 2.4.1.1 Bode-plot

The *Bode-plot*, named after the American engineer Hendrik Wade Bode (1905 – 1982), visualises the frequency response of the output of a dynamic system to an input stimulus and is originally known from analogue electronic signal analysis. Figure 2.24 shows an example of a Bode-plot of a fourth order coupled mass spring system as will be introduced in Chapter 3 on dynamics.

In principle a Bode-plot consists of two parts, one above the other, that both share the same horizontal axis with the frequency as parameter. The upper part is the magnitude Bode-plot with the ratio between the magnitude of the response and the magnitude of the stimulus on the vertical axis. The second part shows the phase shift on the vertical axis relative to phase of the stimulus. Several important aspects have to be recognised in a Bode-plot:

- In a typical Bode-plot both the frequency and the amplitude scales are logarithmic to base 10.
- The frequency scale can be both in rad/s or in Hz. The latter is often preferred in the mechatronic field, mainly because of the more “natural” character of the temporal frequency.
- The magnitude scale can be different for each application because the units of the stimulus and the response can be different. For instance the stimulus can be a force [N] while the response can be a position [m]. The numbers have only a meaning when this relation is known.
- The phase scale can in theory be in radians or degrees. In practice always degrees is used because it enables an easier refinement than with



**Figure 2.24:** Example Bode-plot of a motion system. The upper graph gives the magnitude of the position response divided by the magnitude of the force stimulus. The lower graph gives the phase relation between the position response and the force stimulus. Starting at the low frequency side first the response is frequency independent but at 15 Hz a damped resonance determines the start of a negative slope where the magnitude decreases as function of the frequency. Around 3 kHz a typical resonance and anti resonance indicate the effect of a decoupling mass as will be explained in Chapter 3.

radians while an amount of several degrees can already be important in a control system.

- The Bode-plot represents the response of the system **when all frequencies are continuously present** at the stimulus. It is a stationary representation.

The *double logarithmic* scales are used because most frequency responses are either proportional or inversely proportional to the frequency  $f$  or  $\omega$  to the power  $n$ , where  $n$  is an integer. With double logarithmic scales the function  $\log \omega^n = n \log \omega$  becomes a straight line for all values of  $n$ , because one scale shows  $\log \omega^n$  and the other  $\log \omega$ . The resulting line will have an upward slope when  $n$  is positive and a downward slope when  $n$  is negative. It is common to talk about the order of the slope in terms of  $n$ . For example a transfer function  $f(\omega) = 1/\omega^2$ , which means that  $n = -2$ , results in a line with a slope of  $-2$  in the Bode-plot.

In Section 2.3.2, where the Laplace transform was introduced, a +1 slope was shown to be representative for a differentiating action with a corresponding phase shift of  $+90^\circ$  and a -1 slope is representative for an integrating action with a corresponding phase shift of  $-90^\circ$ . This direct link between phase and slope might lead to the conclusion that the phase plot gives redundant information. In simple “minimum phase” linear systems, the relation between the phase and amplitude is indeed unambiguously determined.

Often, however, this relation is more complicated and the same phase might occur at different frequencies and magnitude-slope values, especially with higher order systems. The influence of sampling, quantisation, hysteresis, backlash and other complicating factors is another reason to not leave the phase plot out in the analysis of the *open-loop* response of modern feedback controlled mechatronic systems. The open-loop phase plot is often even more important for dynamic system analysis than the amplitude. This is the main reason that the Nyquist plot has been developed. Nevertheless, due to its clear relation with the standard applied integration and differentiation operations in motion control systems, the Bode-plot is a frequently used graphical representation both in practice and in this book.

### Absolute magnitude or deci-Bel

The earliest research on frequency behaviour of systems was about the properties of sound. For instance the English physicist John William Strutt (1842 – 1919), better known as Lord Rayleigh, wrote two books on this subject, “The theory of sound Vol 1 and 2” [1], that contains most of the basic theory on dynamics, waves and frequency dependent transfer functions, that still are taught today, including much of the theory in this book.

Sound levels are primarily given in *deci-Bel* (dB). This term originally was introduced by the American engineer Alexander Graham Bell (1847 – 1922) to define the weakening of a signal over a long telephone line. He decided to use a logarithmic scale, because the effect was perceived as a factor per unit of length. As the human perception of sound intensity is also logarithmic, a certain relative change of sound level is always perceived the same, independent of the actual sound level before the change.

Like many signals, sound levels are expressed either in intensity (Power) or a signal amplitude, which is pressure or velocity for sound. These variables can be used both as they are directly interrelated. With sound the power is the multiplication of the complex pressure and velocity.

The  $^{10}$ logarithm of the ratio between two levels of Power is expressed in “Bel” but for reasons of practicality one tenth of this value, the deci-Bel, is chosen

**Table 2.2:** The relation between the order of the slope of a Bode-plot and the magnitude ratio both in deci-Bel and in amplitude and power ratios with a frequency difference of a factor two or ten.

<b>Slope</b>	<b>Per octave (<math>f_1 = 2f_2</math>)</b>			<b>Per decade (<math>f_1 = 10f_2</math>)</b>		
	<b>dB</b>	<b>Amplitude</b>	<b>Power</b>	<b>dB</b>	<b>Amplitude</b>	<b>Power</b>
-3	-18	0.125	0.015625	-60	$10^{-3}$	$10^{-6}$
-2	-12	0.25	0.0625	-40	$10^{-2}$	$10^{-4}$
-1	-6	0.5	0.25	-20	0.1	$10^{-2}$
0	0	1	1	0	1	1
1	+6	2	4	+20	10	$10^2$
2	+12	4	16	+40	$10^2$	$10^4$
3	+18	8	64	+60	$10^3$	$10^6$

as incremental unit.

$$X = 10 \log \frac{P_1}{P_2} \text{ [dB]} \quad (2.46)$$

In electricity the power in a circuit is directly determined by the law of Ohm ( $V = I * R$ ):

$$P = VI = I^2R = \frac{V^2}{R} \text{ [W]} \quad (2.47)$$

For that reason the relation in dB between two magnitude levels of electrical signals is:

$$X = 20 \log \frac{I_1}{I_2} \text{ or } X = 20 \log \frac{V_1}{V_2} \text{ [dB]} \quad (2.48)$$

This means that in expressing the relation between different levels of power, the multiplication factor is ten and between different levels of signal amplitude the multiplication factor is twenty.

In terms of deci-Bel, the -1,-2,+1,+2 etc term for the slope in the Bode-plot becomes equal to a certain number of dB per frequency ratio as shown in Table 2.2. The values are given both for a decade, representing a factor ten between the frequencies, and for an octave with a factor two difference, as commonly used in sound systems and music.

While the dB is frequently used for the magnitude scale of Bode-plots, probably because it avoids the “power of ten” term with large numbers, it is more practical to use the absolute number when analysing the dynamic properties of mechatronic systems. The reason for this preference is that the

magnitude level is the reference for the magnitude of the different control parameters.

In practice the use of dB can even cause confusion when analysing the dynamic properties of a system at different frequencies at a ratio that is neither a factor two or ten. With for instance a factor three frequency difference one might think this is equal to 0.33 of a decade and conclude that a -2 slope (40 dB per decade) would then give an attenuation of  $-0.33 \cdot 40 = -13.2$  dB. In reality the magnitude attenuation would be  $20 \log 1/3^2 = -19$  dB. This error is often made by students at examinations.

In control terminology the *0 dB level* is often used to indicate the frequency where the magnitude of the transfer function of a system becomes equal to one. In case one prefers to avoid the use of dB, like in parts of this book, this frequency can be named the *unity-gain cross-over frequency*.

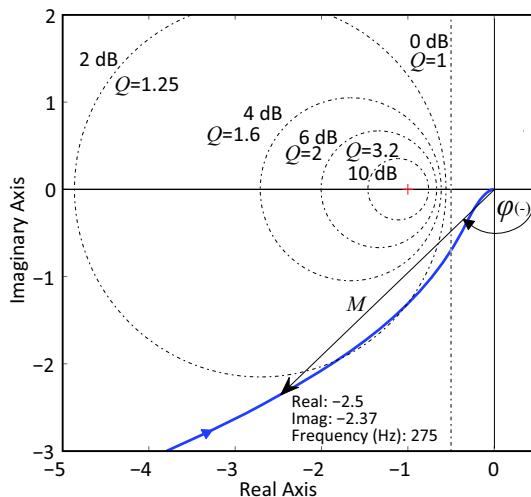
With the Bode-plots in this book the horizontal scale is always in Hz and the magnitude scale is mostly absolute with the exception of the electronics chapter where both notations are used at different sides of the Bode-plot. The main reason for this exception is the possible application of this book by audio engineers for whom it is more logical to use the dB notation and electronic filters are frequently used in their field.

Even though the absolute number is the preferred notation with active control of dynamic systems, also in this field the dB notation is still frequently used. For that reason a mechatronic design engineer should always be aware of these different notations in spite of the apparent logic behind using absolute numbers.

#### 2.4.1.2 Nyquist plot

The *Nyquist plot*, named after the Swedish electrical engineer Harry Nyquist (1889 – 1976), is a polar coordinated two dimensional vector plot where the magnitude  $M$  is represented by the length of the vector starting in the origin. The phase angle ( $\varphi$ ) is represented by the angle of the vector relative to the positive horizontal axis, while its sign is positive moving against the direction of the clock. Because most dynamic systems have a response with a phase delay, most Nyquist plots show a phase angle in the direction of the clock, corresponding with the negative phase that is caused by the delay.

When the different vectors for all frequencies are connected, a curved line is obtained with the frequency as parameter along the line. This plot is often used for the stability and robustness analysis of feedback controlled systems as will be presented in depth in Chapter 4 on motion control. A typical



**Figure 2.25:** Nyquist plot of the open-loop transfer function of a controlled mass positioning system. The blue line connects the polar vector points (black arrow) defined by the phase angle relative to the positive real axis and the magnitude as length of the vector for different frequencies. It starts at low frequencies following the blue arrow to the origin. The phase angle of this example is negative. The Nyquist plot is used to show the stability of a system when the feedback loop is closed.

Nyquist plot of the open-loop response of a simple closed-loop feedback controlled system is shown in Figure 2.25. The blue line represents the response as function of increasing frequency in the direction of the origin. In principle all dynamic systems are not able to respond to extremely high frequencies. As a consequence all Nyquist plots of dynamic systems end in the origin. The distance of the blue line to the dotted circles determine how strongly the closed-loop feedback system will show a resonance. In the shown example the blue line touches the circle of  $2 \text{ dB}/Q = 1.25$  which means that with the same amplitude of the stimulus after closing the loop, the system will show approximately a 25 % larger amplitude of the response at its resonant frequency than at lower frequencies.

# Chapter 3

# Dynamics of motion systems

## Introduction

All mechatronic systems in the context of this book represent controlled motion systems. This statement implies, that they are all inherently dynamic in nature. The dynamic properties of the separate parts within mechatronic systems each have their own impact on the performance with respect to response speed and precision. It is fully justified to say, that the mechanical part is often in practice the most determining factor for the final performance of the total mechatronic system. For this reason it is more than anything else of paramount importance to fully understand the physical properties, that are linked with items like mutually connected springs, bodies and dampers.

The chapter starts with the relation between the stiffness of a mechanical construction and its inherent relative position accuracy under external loads. This will show the relation between accuracy and natural frequency and numerical examples will give a feeling on practical design constraints. It includes an example of an active feedback controlled positioning system as used in a CD-player, in order to link the feedback mechanism with stiffness, as a short introduction towards Chapter 4 on motion control.

This introduction is followed by a more theoretical analysis of the response to a force of a mass-spring<sup>1</sup> system with different levels of damping, the *com-*

---

<sup>1</sup>The term “mass-spring” or “mass-spring-damper” system is a regularly used term although

*pliance* of a system to a force. Closely related is the reaction to a movement of the support, called *transmissibility*. Ample use will be made of Bode-plots and a simplified graphical way to visualise the compliance of each element to further clarify their impact on the overall frequency response. Also a first step towards “multi-body dynamics” will be shown in a coupled mass-spring system. The last section will introduce mode-shapes and the way how the location of the measurement and actuation influences the dynamic transfer function of a mechatronic system.

For good reasons, the dynamic properties presented in this chapter are mostly limited to movements in one direction with linearised properties of the elements. Main argument is, that the use of vector and matrix equations too often diverts the attention from the real physical understanding of the phenomena in relation to the used mathematical models. This basic understanding is often perceived as difficult, while it is essential for mechatronic system design, even more than the ability to master linear algebra as in most cases the computer will do that work for us.

---

a spring and damper are objects, while the mass is a property of a body. To avoid confusion, the regular naming will be used for the complete system, while the element will be named a body with mass  $m$  as its property.

## 3.1 Stiffness

Most people have a rather good qualitative feeling about what is stiff and what is not, but it becomes different when it is necessary to work with quantitative data based on the SI units. The unit of stiffness is called “spring constant”  $k$  and is defined in Hooke’s law of elasticity by the English scientist Robert Hooke (1635 – 1703). It equals the incremental change in the amount of force ( $dF$ ), that an elastic element would produce in reaction to an incremental deformation ( $dx$ ). When the incremental deformation is pointing in the positive  $x$  direction the corresponding incremental force in that same direction<sup>2</sup> would equal:

$$dF = -k(x)dx \implies k(x) = -\frac{dF}{dx} \text{ N/m} \quad (3.1)$$

The negative sign is the result of the fact, that the force is directed in the opposite direction of the deformation, indicating that  $F$  is negative for a positive  $dx$  resulting in a positive number for  $k(x)$ .

With small deformations the stiffness is mostly independent of the deformation. In that case Hooke’s law becomes linear and can be written in the simple notation:

$$F = -kx \implies k = -\frac{F}{x} \quad (3.2)$$

where  $x$  is defined as the displacement from the position where no force is exerted on the spring.

Based on the third law of the English scientist Sir Isaac Newton (1643 – 1727), stating, that the forces of interaction between two bodies are equal, opposite and collinear, an equal force  $F = kx$  must be applied in the direction of the deformation to the elastic element. This equation is also called Hooke’s law, which can cause confusion when it is not clearly defined which force is meant, the action or the reaction. For this reason in this book the second version of Hooke’s law is called the “Hooke – Newton” law even though these two eminent scientists would probably not have liked to be named in one term, in view of their not too well personal relation.

In general the term “spring constant” is associated with the spring as a separate item. The term stiffness is preferred in the technical context

---

<sup>2</sup>The force in all drawings in this book is noted with a double arrow because it acts between two bodies according to the third law of Newton. The direction of the force at the point of insertion is usually defined to be positive in the positive direction of the corresponding position coordinate system defined by one or more arrows, depending on the number of directions in the coordinate system that are presented.

because any object, whether it is a spring or a seemingly solid block has a certain stiffness representing its resistance to deformation. In order to get a feeling for the order of magnitude of the stiffness values in practical mechanical systems it is useful to imagine some real objects and estimate how much they would deform under a known load.

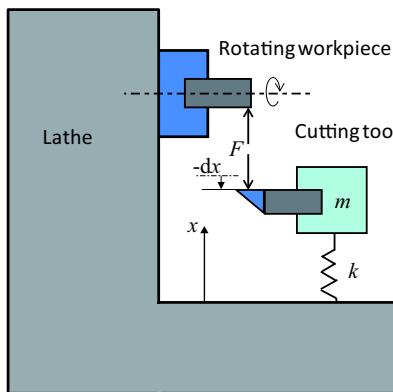
Take for instance a spring board in a swimming pool. Assume, that the mass of the swimmer is 75 kg, exerting 750 N of force and the board moves 10 cm. This corresponds with a stiffness value of the board of  $7.5 \cdot 10^3$  N/m. This is a low value for stiffness, corresponding with the flexible character of a spring board. As a second example in the class-room a load of 1000 kg from 15 students is standing on a strong table. This would represent a force of approximately  $10^4$  N while the table will probably deflect less than 10 mm resulting in a stiffness of about  $10^6$  N/m, which is already a much higher value than with the spring board. An example of a very high stiffness is a steel rod of 20 cm long and one cm diameter, that has a stiffness of  $\approx 10^8$  N/m.

With these and other examples it can be estimated, that practical stiffness values range between approximately  $10^3$  and  $10^9$  N/m, which differ more than six orders of magnitude.

### 3.1.1 Importance of stiffness for precision

To achieve a certain level of accuracy, the stiffness of a system is a very important property. A system with a high stiffness will deform less in response to an applied force and in most cases that is a benefit for a precision system. Examples of systems, that need a reduced stiffness, are for instance a table with four legs and a car with four wheels. Both are in principle *over-constrained* systems meaning that some directions are constrained more than once. Due to the relative flexibility of the tabletop, a table is generally able to deform a bit, making it possible to adapt to floor irregularities and prevent wobbling. With a well designed car suspension system the problem of over-constrain is also solved, while simultaneously increasing the comfort in the car.

An example of a system, where an increased higher stiffness results directly in a better system performance, is the turning lathe that is very schematically shown in Figure 3.1. Depending on the stiffness of the support, the tip of the cutting tool will move away from the intended position due to the cutting forces. The force and the corresponding movement is seldom constant and accurately known so it will limit the possibility to manufacture precise parts.



**Figure 3.1:** Schematic view of a turning lathe. The force that is exerted by the rotating work piece in the negative direction on the cutting tool will deform the spring with stiffness  $k$  resulting in a displacement  $-dx$  from the intended position of the tool tip.

The typical stiffness of a turning lathe is about  $10^7 - 10^8$  N/m. When a maximum error of one micrometre is specified, the change in the cutting force is not allowed to be larger than ten Newton.

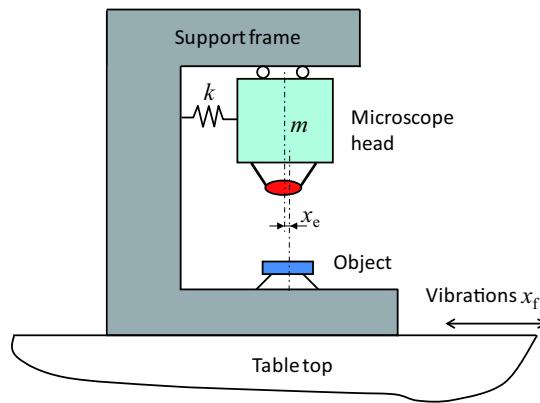
Another example, where high precision is required, is the inspection microscope shown in Figure 3.2. In this example, the desired maximum measurement error is 10 nm. This implies, that the maximum deviation  $x_{\max}$  from the ideal position between the microscope and the object, has to remain smaller than that value. This requirement directly results in a minimum value for the stiffness between the object and the microscope, because of the influence of vibrations of the table top that the microscope is placed on.

The following assumptions are chosen for this example:

The upper part of the microscope has a mass of 5 kg and it is used in a laboratory, where the measured vibrations on the table top show acceleration levels up to  $0.01 \text{ m/s}^2$ . These accelerations have to be followed by the mass of the upper part of the microscope which requires a certain force  $F(t)$  that will cause a deformation of the stiffness  $k$  that is not allowed to be larger than the maximum error of  $\hat{x}_e = 10 \text{ nm}$ . The required stiffness ( $k$ ), to achieve this maximum position error under these circumstances, can be calculated using the second law of Newton:

$$k \geq \frac{F(t)}{\hat{x}_e(t)} = \frac{m \frac{d^2x}{dt^2}}{\hat{x}_e} = \frac{5 \cdot 0.01}{10 \cdot 10^{-9}} = 5 \cdot 10^6 \text{ N/m} \quad (3.3)$$

This value is directly related to the undamped natural frequency  $\omega_0$ , as



**Figure 3.2:** Schematic view of an inspection microscope. Vibrations in the table top act on the entire instrument. Due to the mass and the limited stiffness the microscope head will not be able to perfectly follow the accelerations causing an alignment error  $x_e$  with the object.

determined from the equations of motion according to the following reasoning. A system, that consists of a body and a spring, will resonate in its natural frequency, when the forces inside the system are in balance and the motion is sustained without external forces. This means, that the force  $F_d(t)$ , resulting from the deformation of the spring, is always in equilibrium with the force  $F_a(t)$ , that corresponds with the acceleration of the body:

$$F_a(t) + F_d(t) = m \frac{d^2x}{dt^2} + kx = 0, \quad \Rightarrow m \frac{d^2x}{dt^2} = -kx \quad (3.4)$$

The sign can be checked by the reasoning, that a deformation of the spring in the positive  $x$  direction causes a force and a corresponding acceleration in the opposite direction of the deformation. It is known, that in resonance the body follows a sinusoidal movement  $x(t) = \hat{x} \sin(\omega_0 t)$  with  $\hat{x}$  being the amplitude.

With this information the following equation can be formulated:

$$-m\hat{x}\omega_0^2 \sin(\omega_0 t) = -k\hat{x} \sin(\omega_0 t) \quad (3.5)$$

From which follows:

$$\omega_0 = \sqrt{\frac{k}{m}} \quad (3.6)$$

From the calculated value of the required stiffness in Equation (3.3), the related natural frequency of the system can be derived:

$$\omega_0 \geq \sqrt{\frac{5 \cdot 10^6}{5}} = 1000 \text{ [rad/s]}, \text{ corresponding with:}$$

$$f_0 \geq \frac{1}{2\pi} 1000 = 160 \text{ [Hz]} \quad (3.7)$$

From Equation (3.3) it can also be concluded, that a larger mass of the upper part of the microscope requires a proportionally increased stiffness, to achieve the same accuracy. This would mean that the natural frequency is kept constant. It appears, that the natural frequency is directly related to the performance with respect to the suppression of external vibration disturbances. To show this in mathematical form, the behaviour of the upper part is observed. This microscope head has to follow the sinusoidal motion of the table top vibrations with position  $x(t)$  and acceleration  $\ddot{x}(t)$  equal to:

$$x(t) = \hat{x}_f \sin(\omega t) \implies \ddot{x}(t) = -\hat{x}_f \omega^2 \sin(\omega t) \quad (3.8)$$

The amplitude of the force  $\hat{F}$  acting on the following measurement head is determined by the second law of Newton:

$$F(t) = m\ddot{x}(t) \implies \hat{F} = m\hat{x}_f \omega^2 \quad (3.9)$$

The resulting maximum deformation of the connecting spring equals:

$$\hat{x}_e = \frac{\hat{F}}{k} = \frac{m\hat{x}_f \omega^2}{k} \text{ [m]} \quad (3.10)$$

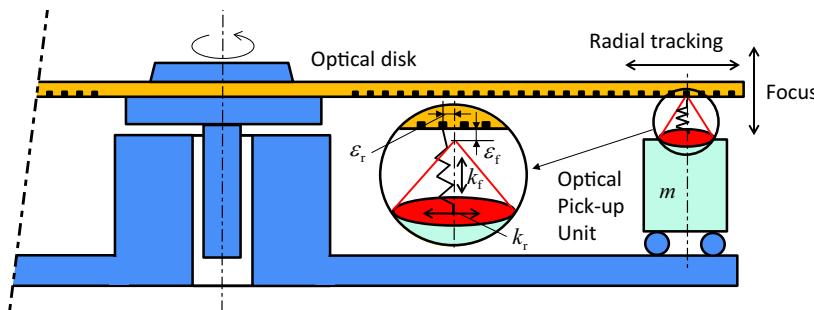
With equation (3.6)  $k/m$  can be replaced by  $\omega_0^2$ , resulting in the following expression:

$$\hat{x}_e = \hat{x}_f \frac{\omega^2}{\omega_0^2} \quad (3.11)$$

The minimal value of  $\omega_0$  to achieve a certain precision becomes:

$$\omega_0 \geq \omega \sqrt{\frac{\hat{x}_f}{\hat{x}_e}} \text{ [rad/s], and } f_0 \geq f \sqrt{\frac{\hat{x}_f}{\hat{x}_e}} \text{ [Hz]} \quad (3.12)$$

This means, that the minimum natural frequency is proportional to the frequency of excitation and to the square root of the ratio between the excitation amplitude and the allowable error. In the next section this same relation is shown in the performance of a feedback controlled positioning system defined by the maximum tracking frequency, called *bandwidth* and the position error  $\epsilon$ . It will show, that these values are related to this natural frequency in passive systems and is a preview to Chapter 4 on motion control.



**Figure 3.3:** Schematic view of an Optical disk system with a “virtual” spring, created by the position control system that connects the pick-up unit in six orthogonal coordinate directions (translations and rotations) to the track. From these directions two translation stiffness values are shown, the radial stiffness  $k_r$  and the focal stiffness  $k_f$  that, together with the mass of the pick-up unit, determine the tracking errors  $\epsilon_r$  and  $\epsilon_f$  under the impact of eccentricity of the disk and external vibrations.

### 3.1.2 Active stiffness

In Chapter 1 the development of the optical disk has been mentioned as an important driver for mechatronics. It required an actively controlled, contact-less positioning of the optical pick-up unit. This example is very appropriate to link precision to the active motion control of Chapter 4. The basic specifications of such an optical disc system can be derived from the requirements. In Figure 3.3 a schematic view is shown. For this example only the radial tracking error ( $\epsilon_r$ ) will be investigated, while the same reasoning can be followed to determine the required stiffness for the focal tracking error ( $\epsilon_f$ ) and the other movement directions including the rotations. For faultless reading of the data, the radial tracking error of the optical pick-up unit has to be smaller than  $0.2 \mu\text{m}$  and the focus error needs to be smaller than  $1 \mu\text{m}$ . The disturbances acting on the system are the eccentricity of the optical disk, that causes a periodic motion of  $200 \mu\text{m}$  in the radial direction at a frequency of  $10 \text{ Hz}$ , and shocks that cause random movements of  $200 \mu\text{m}$  in all directions with the main frequency component at  $25 \text{ Hz}$ .

As a first step, the control system is approximated as if the optical pick-up unit was connected with a spring of a certain radial stiffness  $k_r$  to the track. Like with the previous example, the required stiffness is calculated, based on the accelerations that have to be followed. The largest disturbing force is taken as basis for the calculations. In this case that are the forces, that are caused by the movements due to the shocks, because they have the highest

frequency with the same amplitude as the eccentricity. The forces due to the eccentricity can be neglected, because of the quadratic relation between acceleration amplitude and frequency.

For the calculation it is further assumed, that the movements caused by the shocks can be approximated by a sinusoidal shape. The maximum acceleration of these movements is then described by:

$$\ddot{x}_{\max}(t) = \hat{x}\omega^2 \sin(\omega t) = \hat{x}(2\pi f)^2 = 200 \cdot 10^{-6} (2 \cdot \pi \cdot 25)^2 \approx 5 \text{ [m/s}^2\text{]} \quad (3.13)$$

When the mass of the pick-up unit is  $10^{-2}$  kg, the amplitude  $\hat{F}$  of the disturbing force becomes 0.05 N and with this number and the required maximum error  $\varepsilon_r$  of 0.2  $\mu\text{m}$  the minimum radial stiffness  $k_r$  can be calculated:

$$k_r \geq \frac{\hat{F}}{\varepsilon_r} = \frac{0.05}{0.2 \cdot 10^{-6}} = 2.5 \cdot 10^5 \text{ [N/m]} \quad (3.14)$$

This would result in a natural frequency of:

$$f_0 = \frac{\omega_0}{2\pi} = \frac{1}{2\pi} \sqrt{\frac{k}{m}} = \frac{1}{2\pi} \sqrt{\frac{2.5 \cdot 10^5}{0.01}} \approx 800 \text{ [Hz]} \quad (3.15)$$

When checking this frequency with equation (3.12), the same result is obtained:

$$f_0 \geq f \sqrt{\frac{\hat{x}}{\varepsilon_r}} = 25 \sqrt{\frac{200 \cdot 10^{-6}}{0.2 \cdot 10^{-6}}} \approx 800 \text{ [Hz]} \quad (3.16)$$

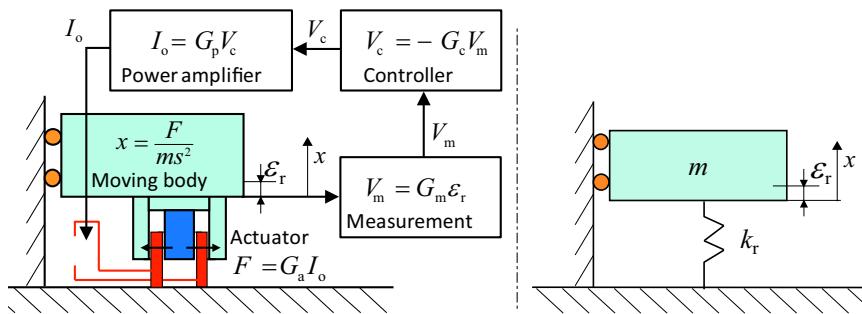
This frequency is called the *roll-off frequency* or *bandwidth* of the controlled dynamic system because above this frequency the system is no longer capable to follow the movements of the track.

In a real optical disk player there is no mechanical contact between the disk and the lens unit, so there is also no real spring. The stiffness has to be caused by something else. In a mechatronic system, the controller takes care of this stiffness. For the CD-player this is a position controller, where the deviation from the set-point position  $\varepsilon_r$  is measured as shown in Figure 3.4. In the shown example of a feedback controller, the error is translated into a force  $F$  from an actuator, opposing to the deviation (negative sign!) with a total gain  $G_t$  equal to the required spring stiffness  $k_r$ :

$$F = G_t \varepsilon_r \Rightarrow G_t = k_r = \frac{F}{\varepsilon_r} \quad (3.17)$$

The total proportional loop gain  $G_t$  consists of the gain of four elements:

- The measurement system, that translates the deviation from the set point into a voltage signal,  $V_m = G_m \varepsilon_r$ .



**Figure 3.4:** Virtual spring created by negative feedback of a deviation  $\varepsilon_r$  of the position from a set point  $x_0$  to a force acting on a body in a servo system. The equivalent spring stiffness equals the total series gain  $k_r = G_t = G_m G_a G_p G_c$  of the feedback loop.

- The control system, that converts this voltage into a negative feedback voltage,  $V_c = -G_c V_m$ .
- The power amplifier that converts this voltage in to a current to the actuator,  $I_o = G_p V_c$ .
- The actuator, that converts this current into a force,  $F = G_a I_o$ .

The negative feedback causes the force to be opposite to the direction of the deviation, just like a passive spring opposes a deformation. This means that a position controller with a constant gain  $G_t$ , acting on a body, creates a virtual spring that behaves fully comparable with a spring in a passive mass-spring system.

As will be shown in the following sections, a simple system with a body and a spring needs damping to return to stand-still after excitation by a stimulus, which is equally true for an active system with only a virtual spring. In Chapter 4 on motion control it will be shown, how a differentiating action can act as damper to stabilise a feedback controlled positioning system.

## 3.2 Mass-spring systems with damping

The mechanics in positioning systems can be modelled as a combination of mass-spring systems, because all material has mass and any structure has a limited stiffness. This somewhat trivial statement has however large implications, as the performance of any positioning system is dominated by the dynamic response of the applied mechanical structure. Two important aspects are considered in the following subsections, the *compliance*<sup>3</sup> of a mechanical system, that describes its dynamic reaction to forces acting on it and the *transmissibility*, that describes the reaction to movements of the supportive structure.

### 3.2.1 Compliance of dynamic elements

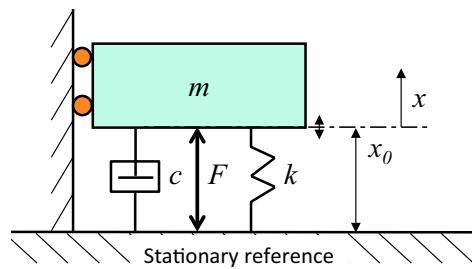
A mass-spring system with damping as shown in Figure 3.5 consists of three elements, that have a different behaviour in respect to forces acting on it. This behaviour is called compliance because it gives a value to the ability of the element to comply, that is to move along with the force. The compliance is given the variable  $C$ .

- The first element, the *spring*, complies with a force by a proportional *displacement*. Its compliance is equal to the inverse of its stiffness  $k$ .  
 $C_s = 1/k$ .
- The second element, the *damper*, complies with a force by a proportional *velocity*. Its compliance is proportional to the inverse of its damping coefficient  $c$ .  
 $C_d \propto 1/c$ .
- The third element, the *body*, complies with a force by a proportional *acceleration*. Its compliance is proportional to the inverse of its mass  $m$ .  
 $C_m \propto 1/m$ .

First the compliance model of a damped mass-spring system is determined, by combining the compliance of the three separate elements in the frequency domain. This helps to create a feeling for the real physics that determine these dynamics. On this base the equations of motion are derived, that show the real response of these damped mass-spring systems.

---

<sup>3</sup>The term “admittance” is also sometimes used for this property. The admittance is more known as the inverse of the impedance in electricity.



**Figure 3.5:** A damped mass-spring system with an external force stimulus. The reaction  $x$  on the force  $F$  is determined by the combined compliance of the body ( $C_m \propto 1/m$ ), the spring ( $C_s \propto 1/k$ ) and the damper ( $C_d \propto 1/c$ ).

### 3.2.2 Combining dynamic elements

In the mass-spring system with damping from Figure 3.5 the applied force is distributed over the three elements while these share the same position  $x$  referenced to the position  $x_0$ , where the system is at rest when  $F = 0$ . Starting with a spring, its compliance  $C_s$  is a simple expression assuming a position independent spring constant:

$$C_s = \frac{x}{F} = \frac{1}{k} \quad [\text{m/N}] \quad (3.18)$$

Note, that the sign is positive, as the displacement is in the same direction as the externally applied force conform the Hooke – Newton law. The compliance of a spring is a constant factor, so it is independent of time and frequency. This means, that an amplitude Bode-plot would show a straight horizontal line with magnitude  $C_s$ . This is called the *spring-line*. In the phase plot the spring-line shows a phase of  $0^\circ$ , because the compliance equation of the spring does not contain any  $(s)$  term in case one would perform a Laplace transform on this constant.

For the damper the force equals in the time domain:

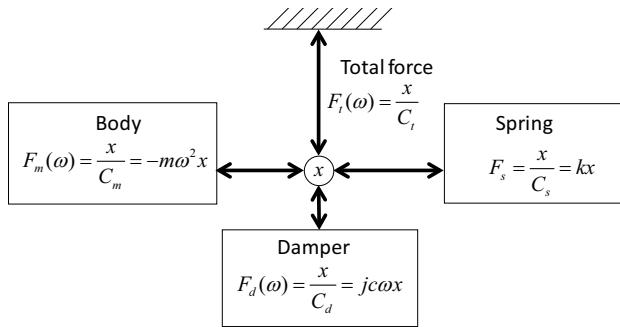
$$F(t) = c \frac{dx}{dt} \quad (3.19)$$

After Laplace transform to bring the function to the frequency domain:

$$F(s) = s cx \implies F(\omega) = jc\omega x \quad (3.20)$$

This results in the following expression for the compliance of the damper as function of the radial frequency ( $\omega$ ):

$$C_d(\omega) = \frac{x}{F} = \frac{1}{jc\omega} \quad [\text{m/N}] \quad (3.21)$$



**Figure 3.6:** In a mass-spring-damper system the force is divided over the three elements according to the inverse of their compliance while sharing the same position.

For the magnitude, the absolute value of the compliance is taken, which means, that the damper has a compliance, inversely proportional to the frequency. In a magnitude Bode-plot this would show up as a straight line with a down slope of  $-1$ , the *damper-line*. This also corresponds with what is understandable by pure reasoning, because the force of a damper is proportional to the velocity and with a harmonic movement  $x = A \sin \omega t$  the related differentiation of the position introduces the additional  $\omega$  term in the amplitude. This means, that with an increase of the frequency the velocity increases if the position amplitude is kept constant. This also means, that a proportional higher force is needed at higher frequencies to move the damper. Special attention should also be given to the  $j$  term in the denominator, which means to say, that the compliance of the damper has an imaginary value and as a consequence the damper-line in the phase plot shows a phase shift of  $-90^\circ$ .

The third element, the body reacts on a force according to the second law of Newton in the time domain:

$$F(t) = m \frac{d^2x}{dt^2} \quad (3.22)$$

After Laplace transform this equation becomes in the frequency domain:

$$F(s) = ms^2 x \quad \Rightarrow \quad F(\omega) = -m\omega^2 x \quad (3.23)$$

This gives the following expression for the compliance of a body as function of the radial frequency ( $\omega$ ):

$$C_m(\omega) = \frac{x}{F} = -\frac{1}{m\omega^2} \quad [\text{m/N}] \quad (3.24)$$

**Table 3.1:** Overview of the dynamic properties of body, spring and damper in the time and frequency domain. The Laplace transform is indicated by a  $\triangleright\triangleleft$  sign and the variable terms ( $t$ ), ( $s$ ) and ( $\omega$ ) are omitted for reasons of simplification.

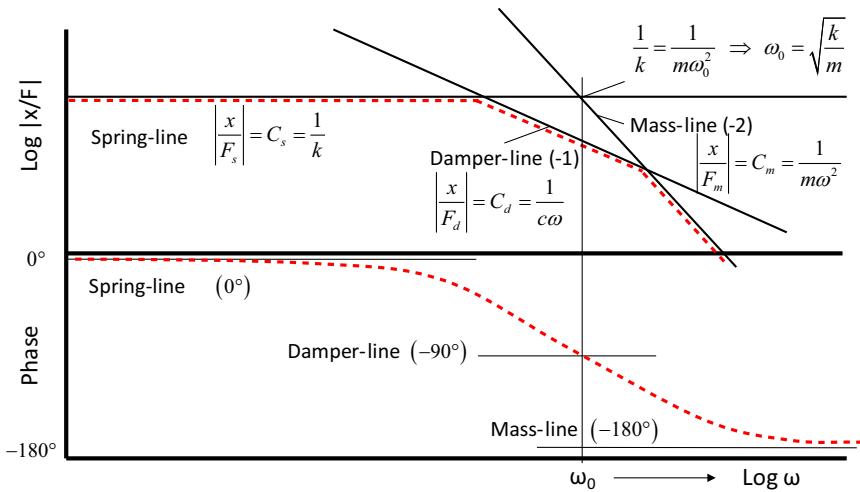
Item	Spring	Damper	Body
<b>Unit</b>	$k$ [N/m]	$c$ [Ns/m]	$m$ [kg]
<b>External force</b>	$F_s = kx$	$F_d = c \frac{dx}{dt} \triangleright\triangleleft csx = jc\omega x$	$F_m = m \frac{d^2x}{dt^2} \triangleright\triangleleft ms^2x = -m\omega^2x$
<b>Compliance [m/N]</b>	$C_s = \frac{x}{F_s} = \frac{1}{k}$	$C_d = \frac{x}{F_d} = \frac{1}{cs} = \frac{1}{jc\omega}$	$C_m = \frac{x}{F_m} = \frac{1}{ms^2} = -\frac{1}{m\omega^2}$
<b>Magnitude [m/N], Phase angle [°]</b>	$\frac{1}{k}, 0^\circ$	$\frac{1}{c\omega}, -90^\circ$	$\frac{1}{m\omega^2}, -180^\circ$

The compliance of the body appears to be proportional to the inverse of the frequency squared. In the amplitude Bode-plot this is represented by a straight line with a slope of  $-2$ . This line should have been called the Body line but, like with the naming of the mass-spring system this line is called the mass-line. The  $-2$  slope is also understandable by reasoning because with a harmonic movement  $x = A \sin \omega t$  acceleration introduces a term  $\omega^2$  in the amplitude of the acceleration, due to the double derivative over time. The minus sign is the result of  $j^2 = -1$  in the denominator, which means, that the phase plot of the mass-line shows a phase shift of  $-180^\circ$ .

In Table 3.1 these relations are put together as a reference. When the three elements are combined, such that they share the same position, the total excitation force will be divided over the three elements, as shown schematically in Figure 3.6. At first sight it is to be expected, that at any frequency the total compliance  $C_t = x/F$  can never exceed the level of the compliance of each of the elements. Although this assumption is not fully correct, as will be shown later, it is useful to start with, in order to get a feel for the relations.

With this assumption the following can be stated:

$$F_t = F_s + F_d + F_m = x \left( \frac{1}{C_s} + \frac{1}{C_d} + \frac{1}{C_m} \right) = \frac{x}{C_t} \quad (3.25)$$



**Figure 3.7:** The response of body, damper and spring, each represented as separate elements by a straight line in a Bode-plot with amplitude and phase. The red dashed line shows the combined compliance plot. The natural frequency  $\omega_0$  where the response will roll-off is found at the intersection of the spring and mass magnitude lines.

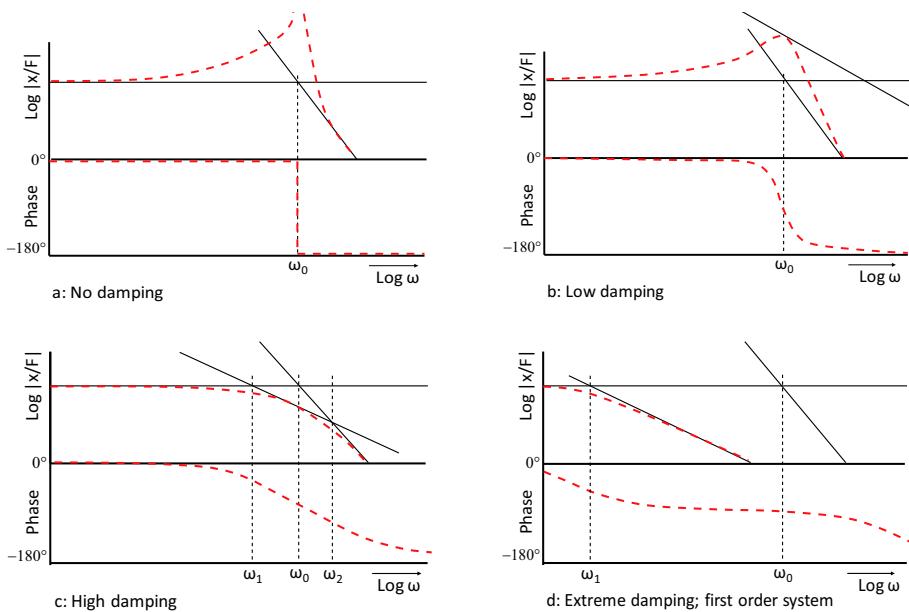
Which means, that the total compliance equals:

$$C_t = \frac{x}{F_t} = \frac{1}{\frac{1}{C_s} + \frac{1}{C_d} + \frac{1}{C_m}} \quad [\text{m/N}] \quad (3.26)$$

This corresponds with the statement, that at any frequency the element with the least compliance determines the total compliance.

To illustrate the individual effect on the Bode-plot of each element, each response is drawn separately in the amplitude and phase part of a Bode-plot as shown in Figure 3.7. In this Plot the frequency is noted in the angular frequency ( $\omega$ ), because of the clear relationship with the mathematical expression for the compliance terms for each element.

The magnitude of the compliance of the combined mass-spring-damper system can now be derived by taking the lowest value of the individual elements depending on the frequency region. At low frequencies, the spring determines the response of the system with the horizontal spring-line at a compliance magnitude-level of  $1/k$ . At high frequencies the behaviour is fully determined by the body, according to its mass-line with a magnitude level of  $1/m\omega^2$ . The *roll-off* starts at the intersection of these lines at the natural frequency  $\omega_0$ . This can be concluded from the following reasoning.



**Figure 3.8:** The Bode-plots of the response of a mass-spring system with different levels of damping shows the effect on the resonance at the natural frequency  $\omega_0$ .

At intersection both magnitudes are equal, so:

$$\frac{1}{k} = \frac{1}{m\omega^2} \quad \Rightarrow \quad \omega^2 = \frac{k}{m} \quad \Rightarrow \quad \omega = \sqrt{\frac{k}{m}} = \omega_0 \quad (3.27)$$

In the mid frequency range, the damper determines the behaviour of the system. Depending on the level of the magnitude ( $1/c\omega$ ) of its corresponding damper-line, the phase shows a gradual combination of the phase shift of the different elements.

To get an initial idea about the effects of different levels of damping four examples are shown in Figure 3.8.

Bode plot 3.8.a shows the response of mass-spring system without damping. This is the situation where the assumption on the upper limitation of the compliance is no longer true. At the natural frequency  $\omega_0$  the amplitude of the system increases by absorbing energy and loading it in alternating potential energy (spring) or kinetic energy (mass), resulting in an infinite compliance. At this natural frequency the phase jumps from 0° to -180° without gradual transition. In the next section this effect will be better explained with the equations of motion.

As soon as a small damper is added, as shown in Bode plot 3.8.b, the peak in the compliance at  $\omega_0$  is suppressed by the damper that absorbs energy out of the dynamic system. Because the damping is still moderately strong, the original infinite response is now only lowered to a finite value. When a damper with a higher damping coefficient is applied, as shown in Bode plot 3.8.c, the damping-line shifts further down in the amplitude plot. If the damping is high enough, such that the damping-line is below the intersection of the spring- and mass-line, the amplitude at the resonance frequency is determined purely by the damping and the resonance peak will not occur. This situation corresponds with a series of two low-pass first-order systems, the first starting at a roll-off frequency  $\omega_1 < \omega_0$  and the second starting at a roll-off frequency  $\omega_2 > \omega_0$ . If the damping is even further increased, the mass-spring system is “over-”damped as shown in Bode plot 3.8.d. The damper-line in the amplitude plot shifts almost completely below the mass-line. In this situation the dynamic response can be approximated by a first-order system with a roll-off frequency of  $\omega_1$ . Above this frequency the damper dominates the system behaviour over the mass while  $\omega_2$  becomes a very high frequency where the magnitude of the response is very low.

### 3.2.3 Transfer functions of the compliance

In this section the modelled frequency response of the damped mass-spring system will be described in the form of a frequency dependent *transfer function* (TF) with the Laplace variable  $s = j\omega$  as variable. While transfer functions work both in the time- and frequency domain, in the latter case they are also often called the *frequency response function* (FRF) of the dynamic system. In this book the general term ”transfer function“ will be used for both domains. These functions are derived from the reaction of a body to a force with equations of motion that describe the position, velocity and acceleration of the body. First the compliance behaviour will be derived, as was schematically presented in the previous section. Although the angular frequency is used in the mathematics of the transfer functions, the temporal frequency will be shown in the graphical representations, because of its clear relation with the engineering field of use. For a mechatronic designer it is important to recognise and be able to work with this “dual” expression with a  $2\pi$  difference.

### 3.2.3.1 Damped mass-spring system.

To mathematically model the compliance of a generic mass-spring system with a damper, the same configuration is used, as shown in Figure 3.5 in the previous section. The system compliance equals the movement ( $x$ ) of the position relative to the position  $x_0$ , where the system is at rest, in response to a force ( $F$ ), exerted between the body and the stationary reference. The object is guided with freedom to move in only one direction by a linear guiding system, represented by the rollers. In the  $x$  direction, where the force is acting, the spring with stiffness  $k$  and the damper with damping coefficient  $c$  impact the motion of the body with mass  $m$  by their respective compliances as described before.

The mathematical analysis begins by taking the balance of forces acting on the body in the time domain:

$$F(t) = m \frac{d^2x}{dt^2} + c \frac{dx}{dt} + kx \quad (3.28)$$

In writing down this *second-order* differential equation it is necessary to carefully look at the sign of the terms. With an external force in the upward (positive) direction, the acceleration, velocity and position will all work in that same direction.

Using the Laplace transform this differential equation can be written in the frequency domain as:

$$F(s) = ms^2x + csx + kx = x(ms^2 + cs + k) \quad (3.29)$$

From this equation the following relation for the total compliance  $C_t(s)$  can be derived:

$$\begin{aligned} C_t(s) &= \frac{x}{F} = \frac{1}{ms^2 + cs + k} \\ &= \frac{1}{\frac{m}{k}s^2 + \frac{c}{k}s + 1} \quad [\text{m/N}] \end{aligned} \quad (3.30)$$

To qualify this expression, three entities are introduced, the damping ratio  $\zeta$ , the spring compliance  $C_s$  and the un-damped<sup>4</sup> natural frequency  $\omega_0$ . Of these the damping ratio is a new term. It is a dimensionless number as

---

<sup>4</sup>The un-damped natural frequency is mostly named just the natural frequency, which is not completely correct as at higher damping values the peak in the compliance response will shift slightly to a lower frequency as will be presented a bit further on in this chapter.

it gives the ratio between the actual damping coefficient  $c$  and the critical damping coefficient  $c_0$  defined by the level where the step response will not show overshoot anymore, as will be shown in Figure 3.12 in the next section. This is the situation when this second-order mass-spring system just behaves like a combination of two equal first-order systems. It will be shown later, that this critical damping coefficient equals  $2\sqrt{km}$ .

To summarize the derived variables:

$$\begin{aligned}\zeta &= \frac{c}{2\sqrt{km}} \\ C_s &= \frac{1}{k} \\ \omega_0 &= \sqrt{\frac{k}{m}}\end{aligned}\tag{3.31}$$

Using these terms in Equation (3.30) gives the following:

$$C_t(s) = \frac{x}{F} = \frac{C_s}{\frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1} \quad [\text{m/N}]\tag{3.32}$$

To determine the amplitude of the signal the Laplace variable  $s$  is substituted by  $j\omega$  to give the equation as function of the radial frequency ( $\omega$ ):

$$C_t(\omega) = \frac{x}{F} = \frac{C_s}{\frac{j^2\omega^2}{\omega_0^2} + j2\zeta \frac{\omega}{\omega_0} + 1} = \frac{C_s}{\underbrace{1 - \frac{\omega^2}{\omega_0^2}}_{\text{Real}} + \underbrace{j2\zeta \frac{\omega}{\omega_0}}_{\text{Imaginary}}} \quad [\text{m/N}]\tag{3.33}$$

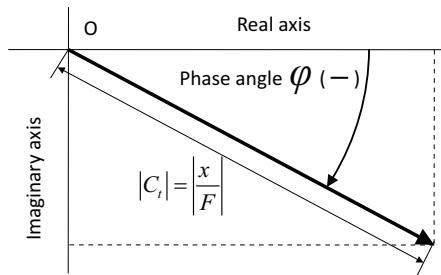
The amplitude is the absolute value of the vector, being the square root of the sum of the squared values of the real part and the imaginary part, as shown in the Nyquist plot of Figure 3.9. This means, that the amplitude equals:

$$|C_t|(\omega) = \left| \frac{x}{F} \right| = \frac{C_s}{\sqrt{\left(1 - \frac{\omega^2}{\omega_0^2}\right)^2 + \left(2\zeta \frac{\omega}{\omega_0}\right)^2}}\tag{3.34}$$

As a first conclusion, in case of no damping ( $\zeta = 0$ ), the response becomes infinite, when  $\omega$  equals  $\omega_0$ . This is the reason, that the frequency  $f_0 = \omega_0/2\pi$  is called the natural or resonance frequency of the mass-spring system.

In presence of damping the maximum value for  $C_t$  at this frequency relative to the spring-line becomes:

$$\frac{|C_t|_{\max}}{C_s} = \frac{1}{2\zeta}\tag{3.35}$$



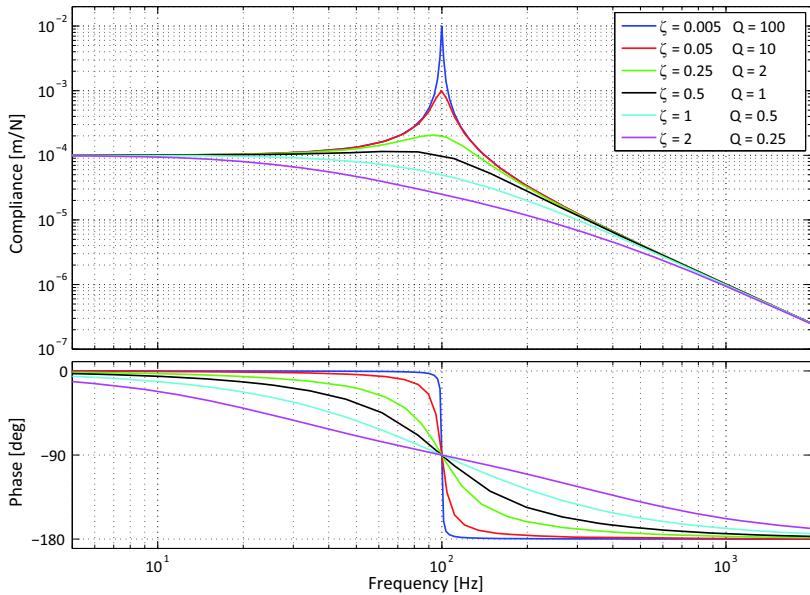
**Figure 3.9:** Polar (Nyquist) representation of the response of a damped mass-spring system at one frequency.

In Chapter 2 it was shown, that the Laplace variable  $s$  introduces a phase shift of  $+90^\circ$ , when located in the numerator of the transfer function. Located in the denominator of the transfer function the Laplace variable  $s$  introduces a phase shift of  $-90^\circ$ . This means, that the phase angle of any dynamic transfer function equals the phase angle of the numerator part minus the phase angle of the denominator part. While the numerator and denominator generally consist of both real and imaginary terms the phase angle  $\phi$  of both numerator and denominator is determined by the arctangent of the ratio of their respective imaginary and real terms where the quadrant of the angle is given by the signs of the terms. For a positive real and imaginary value the angle equals:

$$\text{Phase} = \phi = \arctan\left(\frac{\text{Imaginary}}{\text{Real}}\right) \quad (3.36)$$

The angle for a negative real value and a positive imaginary value equals  $180^\circ$  minus the above calculated arctangent value. A negative value for both gives an angle of  $180^\circ$  plus the above calculated arctangent value while a positive value for the real term and negative value for the imaginary term gives a phase angle of  $360^\circ$  minus the calculated arctangent value.

For this damped mass-spring system the numerator of Equation (3.32) is constant without imaginary part. As a consequence the phase angle of the numerator is zero which means that the phase angle is only determined by the denominator with a negative sign added to the above calculation. This all is summarised in the following equation which works in two quadrants because the imaginary term is always positive and only the real term has two signs, positive at a frequency below  $\omega_0$  and negative at a frequency



**Figure 3.10:** Bode-plot of the compliance of a damped mass-spring system with  $k = 1 \cdot 10^4$  N/m and  $m = 0.025$  kg, giving a natural frequency  $f_0$  of 100 Hz, at different values for  $\zeta$  and  $Q$ .

above  $\omega_0$ :

$$\phi_{\text{tot}} = \phi_{\text{num}} - \phi_{\text{den}} = \arctan\left(\frac{0}{C_t}\right) - \arctan\left(\frac{2\zeta\frac{\omega}{\omega_0}}{1 - \frac{\omega^2}{\omega_0^2}}\right) = -\arctan\left(\frac{2\zeta\frac{\omega}{\omega_0}}{1 - \frac{\omega^2}{\omega_0^2}}\right) \quad (3.37)$$

To show the amplitude and the phase of this transfer graphically, an example damped mass-spring system is defined, that will be used throughout most of this section. The spring has a spring constant  $k$  of  $1 \cdot 10^4$  N/m and the body has a mass  $m$  of 0.025 kg, giving a natural frequency of:

$$f_0 = \frac{1}{2\pi} \sqrt{\frac{k}{m}} = \frac{1}{2\pi} \sqrt{\frac{1 \cdot 10^4}{0.025}} = 100 \quad [\text{Hz}] \quad (3.38)$$

With the mathematical software MATLAB, Equation (3.32) results in the Bode-plot of Figure 3.10. Clearly the same shape of the plot is shown as derived by the graphical method of the previous section. Below  $\omega_0$  the graph starts with the horizontal spring-line, corresponding to a value of  $C_s = 1/k = 1 \cdot 10^{-4}$ . Above  $\omega_0$  the graph follows the mass-line  $1/m\omega^2$  with a slope of  $-2$ . The compliance is calculated for different levels of damping.

The maximum value of the peak at for instance  $\zeta = 0.25$  is not at  $\omega_0$ , but at a slightly lower frequency, which behaviour will be explained in the following section. It is also shown, that Equation (3.35) is correct, as for instance the resonance peak at  $\zeta = 0.005$  has a magnitude of a factor hundred above the magnitude of the spring-line. This factor hundred is equal to the variable  $Q$ , the *quality factor*, that equals  $Q = 1/2\zeta$ . This variable will be further presented later, after the next section.

### 3.2.3.2 Critical damping and definition of $\zeta$

The definition of  $\zeta$  is based on the analysis of the *poles* of the compliance transfer function. Poles are those values of  $s$ , that result in a zero in the denominator of the transfer function, corresponding with an infinite value of the transfer function. This is the most easily understood at the situation of an un-damped mass-spring system where at  $s = \pm j\omega_0$  the magnitude of the compliance becomes infinite. Poles can be shown in a geometric complex plane representation that is called the *Laplace plane* because of the relation with the Laplace variable  $s$ . The poles of a second-order system appear as complex conjugate terms, because of the squared relation, and in the undamped situation the poles are purely located on the imaginary axis, symmetrical around the real axis. For the damped situation the full definition of the Laplace variable  $s = \sigma + j\omega$  must be used, so including the real part  $\sigma$ , to determine the pole location. The denominator ( $d_c$ ) of the compliance transfer function is a second-order differential equation with variable  $s$ . To determine the poles this equation can be written in a more generalised form as a multiplication of two terms:

$$d_c(s) = (s - p_1)(s - p_2) \quad (3.39)$$

where  $p_1$  and  $p_2$  are the two poles. Generally these poles are complex numbers and come as two conjugate complex terms, written as:

$$p_1 = -\sigma + j\omega_d \quad \text{and} \quad p_2 = -\sigma - j\omega_d \quad (3.40)$$

where  $\sigma$  is the real part of the pole and the imaginary part contains  $\omega_d$ , the real resonance frequency in the damped situation. When these are combined in the generalised description of the denominator, it becomes:

$$d_c(s) = (s + \sigma - j\omega_d)(s + \sigma + j\omega_d) = (s + \sigma)^2 + \omega_d^2 \quad (3.41)$$

As a next step, equation (3.30) is written in a different form, in order to come to a denominator notation, that can be compared with this pole notation.

$$C_t(s) = \frac{x}{F} = \frac{\frac{1}{k}}{\frac{m}{k}s^2 + \frac{cs}{k} + 1} \quad (3.42)$$

Introducing the natural frequency  $\omega_0 = \sqrt{k/m}$  and the spring compliance  $C_s$  results in:

$$C_t(s) = \frac{x}{F} = \frac{C_s}{\frac{s^2}{\omega_0^2} + \frac{cs}{k} + 1} \quad (3.43)$$

Multiplication of both the numerator and the denominator of the equation with  $\omega_0^2$  gives:

$$C_t(s) = \frac{x}{F} = \frac{C_s \omega_0^2}{s^2 + \frac{cs \omega_0^2}{k} + \omega_0^2} \quad (3.44)$$

For the pole locations only the denominator part is relevant:

$$d_c(s) = s^2 + \frac{cs \omega_0^2}{k} + \omega_0^2 \quad (3.45)$$

To determine the poles the mid term is changed by replacing one factor  $\omega_0$  by  $\sqrt{k/m}$ :

$$d_c(s) = s^2 + \frac{cs \sqrt{\frac{k}{m}} \omega_0}{k} + \omega_0^2 = s^2 + \frac{cs \omega_0}{\sqrt{km}} + \omega_0^2 \quad (3.46)$$

When the middle and the last term of this equation are compared with the corresponding terms in the expanded version of equation (3.41):

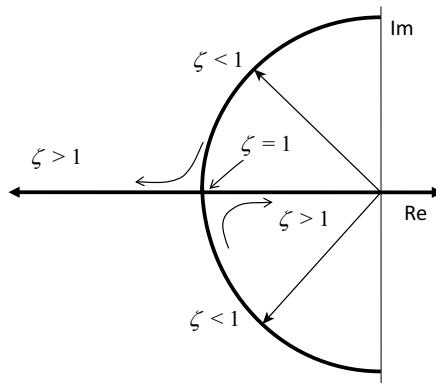
$$d_c(s) = (s + \sigma)^2 + \omega_d^2 = s^2 + 2\sigma s + \sigma^2 + \omega_d^2 \quad (3.47)$$

then the following equations are obtained that relate the real and imaginary value of the pole to  $\omega_0$ :

$$\sigma = \frac{c \omega_0}{2\sqrt{km}} \quad \text{and} \quad \omega_0^2 = \sigma^2 + \omega_d^2 \quad (3.48)$$

The second relation is a circle equation that can be written as follows to obtain the imaginary term  $\omega_d$  of the pole:

$$\omega_d = \sqrt{\omega_0^2 - \sigma^2} = \omega_0 \sqrt{1 - \frac{\sigma^2}{\omega_0^2}} \quad (3.49)$$



**Figure 3.11:** Pole locations of a damped mass-spring system in the complex plane.

Depending on the damping ratio the poles are either real numbers, that represent an overly damped system consisting of two first-order systems, or complex conjugate numbers with less damping. The magnitude of the real part relative to the imaginary part determines the amount of damping in the system.

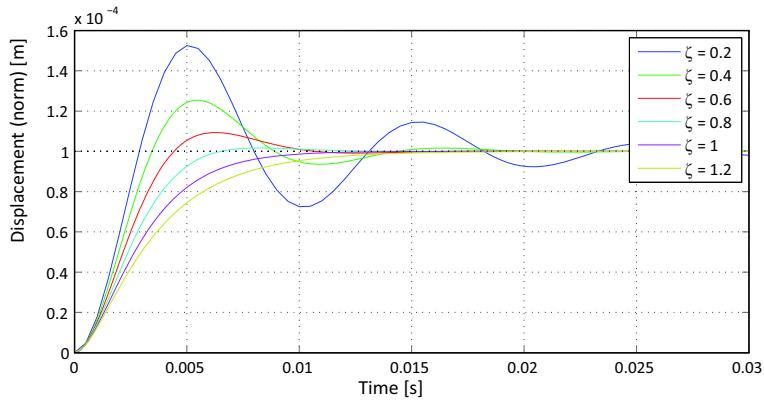
With Equation 3.48 for the relation of  $\sigma$  this results in:

$$\omega_d = \omega_0 \sqrt{1 - \frac{c^2}{4km}} \quad (3.50)$$

In the situation where the term  $c^2/4km$  becomes equal to one, the poles become two equal real negative poles. This means, that the second-order transfer function of the dynamic system is reduced to a combination of two equal *first-order* systems. This situation corresponds with the condition  $c = 2\sqrt{km}$ , that is called the critical damping coefficient with  $\zeta = c/(2\sqrt{km})$  as the related critical damping ratio.

When the damping is increased beyond that level the roll-off frequencies of the two first-order systems become separated as was shown in Section 3.2.2. The lower roll-off frequency corresponds with the pole that shifts to the right in the complex plane and the higher roll-off frequency corresponds with the pole that shifts to the left. The corresponding real poles of these two first-order systems are equal to  $-1/\tau = -\omega_0$  where  $\tau$  equals the time constant of the system and  $\omega_0$  the roll off frequency.

Figure 3.11 shows graphically the effect of the damping on the pole locations in the complex plane for a damped mass-spring system, corresponding to the circle relation of the pole terms in Equation (3.48). The fact, that damping corresponds with a dominant negative real term will be used also



**Figure 3.12:** Position response of a mass-spring system in the time domain with  $k = 1 \cdot 10^4$  N/m and  $m = 0.025$  kg to a step Force of 1 N with different levels of damping.

in Chapter 4 on motion control where a stable system requires the poles to be located in the left half of the complex plane. It will also be shown in that chapter with the presentation of state-space feedback control that these poles correspond to the *eigenvalues* of an *eigendynamics* matrix  $\mathbf{A}$  that contains the terms of the transfer function in a vector-matrix notation.

From the above equations it follows that increasing the damping will shift the real resonance frequency  $\omega_d$  below the un-damped natural frequency  $\omega_0$ , until the real resonance frequency becomes equal to zero at  $\zeta = 1$ . As a direct consequence, the overshoot at a step response is zero, as can be seen in Figure 3.12.

### 3.2.3.3 Quality-factor $Q$

In electrical engineering the “Quality-factor”  $Q$  is defined as a variable for the measure in which a system resonates. The value of  $Q$  is determined by the maximum value of the peak in the resonant system, relative to the level of the spring-line. Electrical engineering is a typical frequency domain oriented discipline and also  $Q$  belongs to that domain, because of the emphasis on the behaviour at the natural frequency. Furthermore, resonance is often positively valued in electronics for tuning filters, stabilising clocks and other useful functions. For this reason  $Q = 1$  is defined as the minimum level where just no resonance occurs anymore.

In mechanical engineering, with its typical time domain orientation, the damping ratio  $\zeta$  is more common to use, because of its time response relation and also because resonance is often seen as a negative effect. When possible, a well behaved step response is aimed for which is the case at  $\zeta = 1$ , as was demonstrated in the previous section. In fact both terms can be used together depending on the situation. In case of a resonator, a high  $Q$  value is preferred, while a controlled, well damped system demands for a high  $\zeta$ . Fortunately there is a very simple and straightforward relation between  $Q$  and  $\zeta$ :

$$Q = \frac{\sqrt{km}}{c} = \frac{1}{2\zeta} \quad (3.51)$$

The direct relation of  $Q$  with the resonance peak in the Bode-plot is shown when combining Equation (3.51) with Equation (3.35):

$$\frac{|C_t|_{\max}}{C_s} = \frac{1}{2\zeta} = Q \quad (3.52)$$

This result was also shown in Figure 3.10. In the example damped mass-spring system with spring constant  $k$  of  $1 \cdot 10^4$  N/m and a body with a mass  $m$  of 0.025 kg, this means, that an excitation force amplitude of 1 N will result in a 0.1 mm motion amplitude at the spring-line, a 1 mm peak amplitude at  $\omega_0$ , when  $Q = 10$ , and a 10 mm peak amplitude at  $\omega_0$ , when  $Q = 100$ .

A special situation occurs when  $Q = \zeta = \sqrt{0.5} \approx 0.7$ . In that case neither the time domain response nor the frequency domain response shows any periodicity for which reason this damping level is called *aperiodic*. It is an optimal situation defined by the shortest response time of a dynamic system without the occurrence of periodic movements due to the first resonance frequency of the system.

$Q$  and  $\zeta$  also have a relation with the energy in the system. In case of a resonating mechanical system the energy from the actuator is stored pe-

riodically in kinetic energy (mass) and potential energy (spring). When the supply of energy is interrupted, the resonance will gradually decrease, because in practice a part of the energy is lost in every cycle due to unavoidable damping effects in the material like hysteresis, air friction, and so on. The mechanism causing this gradual loss of energy is based on the understanding, that the damping force and velocity are in phase and opposite to each other. This means, that power is dissipated by the damper at a rate, equal to the scalar multiplication of force times the velocity:

$$P(t) = F \frac{dx}{dt} = -c \frac{dx}{dt} \frac{dx}{dt} = -c \left( \frac{dx}{dt} \right)^2 = -cv^2 \quad (3.53)$$

This energy is dissipated into heat.

The exact relation of  $Q$  with energy is determined by looking at the ratio between the stored energy in a resonating system and the dissipated or lost energy per cycle. At the natural frequency  $\omega_0$  the stored energy ( $E_s$ ) is equal to the kinetic energy of the body at its maximum speed ( $\hat{v}$ ) when the spring is unloaded:

$$E_s(t) = \frac{1}{2} m \hat{v}^2 \quad (3.54)$$

The lost energy per cycle ( $E_l$ ) at the natural frequency  $\omega_0$  is calculated by using Equation 3.53 with the sinusoidal velocity  $v = v_m \sin \omega_0 t$  and integrating the power over a time equal to the period ( $T$ ), which results in the following:

$$E_l(t) = \int_0^T \hat{v}^2 \sin^2 \omega_0 t dt = \frac{1}{2} c \hat{v}^2 T = \frac{1}{2} c \hat{v}^2 \frac{2\pi}{\omega_0} \quad (3.55)$$

With  $\omega_0 = \sqrt{k/m}$  and using above equations the energy ratio ( $R_e$ ) equals:

$$R_e = \frac{E_s}{E_l} = \frac{\frac{1}{2} m \hat{v}^2}{\frac{1}{2} c \hat{v}^2 \frac{2\pi}{\omega_0}} = \frac{m \omega_0}{2\pi c} = \frac{\sqrt{km}}{2\pi c} \quad (3.56)$$

Because  $Q = 1/2\zeta = \sqrt{km/c}$  the following conclusion is valid:

$$Q = 2\pi R_e = 2\pi \frac{\text{Maximum energy stored}}{\text{Energy lost per cycle}} \quad (3.57)$$

Note, that the mentioned “maximum energy stored” is the energy that is present in the system at each specific cycle and this value decreases continuously as a consequence of the dissipated energy in the damping per cycle.

### 3.2.3.4 Behaviour around the natural frequency

As a thought experiment, at  $t = 0$ , a mass-spring system without damping is excited at its un-damped natural frequency by means of a sinusoidal force with a constant amplitude. As of  $t = 0$  a continuous increase of the amplitude will be observed, depending on the amplitude of the force in relation to the mass of the body. This effect is fully comparable with a proportional linear acceleration of a body in response to a constant force.

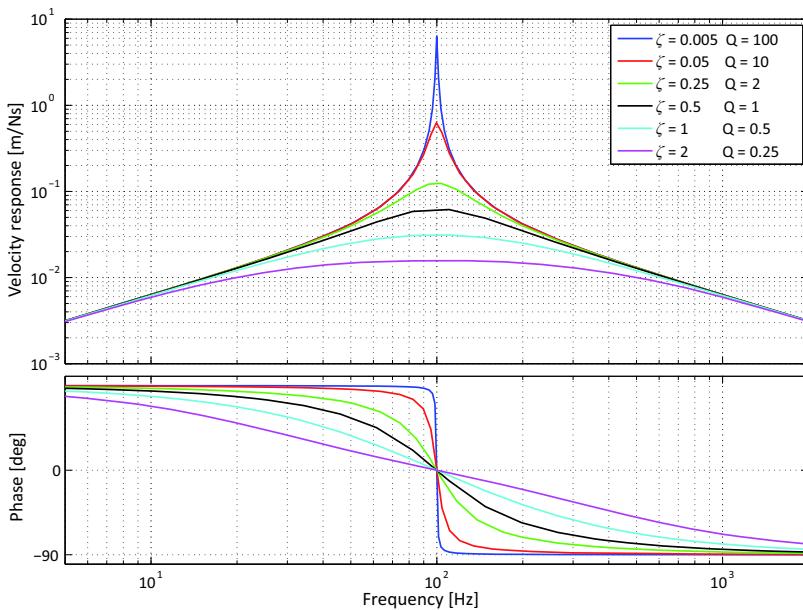
It was mentioned before, that at the the un-damped natural frequency of a mass-spring system, the energy is continuously exchanged between potential energy in the spring and kinetic energy in the body. Due to the ongoing excitation, the captured amount of energy in the system is growing constantly, causing *ultimately* an infinite amplitude of motion.

In practice the excitation frequency is hardly ever completely constant. Further always some damping is present due to hysteresis and friction. Finally springs with an infinite linear strain do not exist either. This all means, that this infinite gain will never occur in reality.

Nevertheless it is extremely important to iterate that a Bode-plot is a stationary representation, based on the presence of a continuous stimulus at all frequencies. This means, that if the excitation frequency is not sufficiently long available at the resonance frequency, the resonating effect will have no time to develop. It also means, that extremely heavy systems with a very low resonance frequency, like for example maritime pontoons floating in the sea, will take a long time before the effect becomes visible, when the waves are exciting the pontoons at their resonance frequency. Because the amount of energy per cycle added to the system by the waves is small, the effect might be only significant or even disastrous after a long period.

This relation with energy is also clear, when observing the phase behaviour. At the natural frequency the phase between force and position is  $-90^\circ$ , while also the velocity has a  $-90^\circ$  phase difference with the position in case of sinusoidal movements. This means, that the force is in phase with the velocity at the natural frequency, thus acting in the direction of the movement and maximising the work. This leads to the conclusion, that in a vibrating system the maximum efficiency in energy transfer from the force of an actuator to a mechanical movement is realised at the natural frequency.

In Figure 3.13 this phenomenon is illustrated by the Bode-plot of the velocity response of the example mass-spring system with spring constant  $k$  of  $1 \cdot 10^4 \text{ N/m}$  and a body with a mass  $m$  of  $0.025 \text{ kg}$  to an external force. This velocity response transfer function is found simply by differentiation over



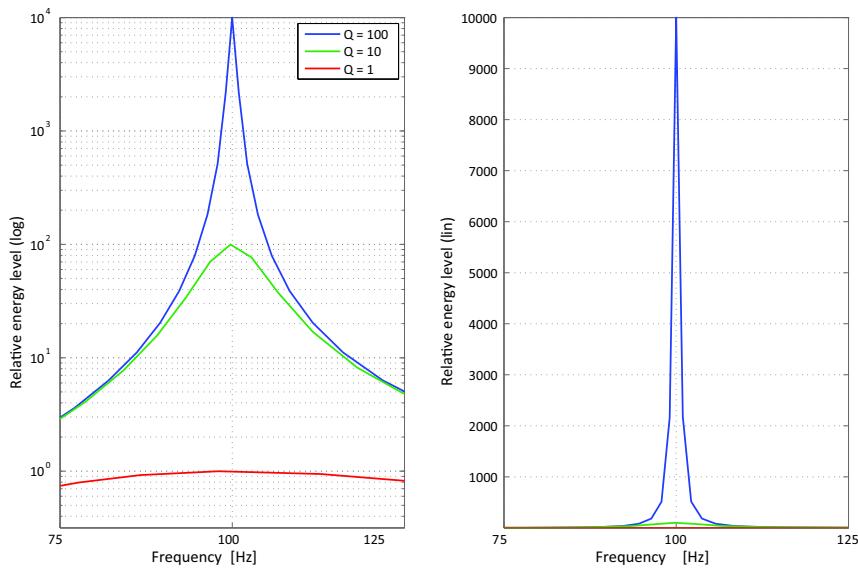
**Figure 3.13:** Bode-plot of the response of the velocity of a damped mass-spring system with  $k = 1 \cdot 10^4$  and  $m = 0.025$  kg to a force with different values for  $\zeta$  and  $Q$ . It clearly shows, that the force is in phase with the velocity at the natural frequency, achieving the maximum efficiency in energy transfer.

time of the time domain transfer function of the position. In the frequency domain this differentiation is equal to the multiplication of the system compliance  $C_t(s)$  with  $s$ :

$$\frac{v}{F}(s) = \frac{sx(s)}{F} = sC_t(s) = \frac{sC_s}{\frac{s^2}{\omega_0^2} + 2\zeta\frac{s}{\omega_0} + 1} \quad (3.58)$$

The peak level at for instance  $Q = 100$  can be checked by taking the earlier found motion amplitude of 10 mm and multiply that value with  $\omega$  to get the maximum velocity. At  $100$  Hz = 628 rad/s this results in a maximum velocity of 6.28 m/s, that corresponds with the level of the modelled peak in the Bode-plot.

As a closure of this part on damping and energy it is illustrative to draw a plot of the velocity squared over a small area around the natural frequency. Because the stored kinetic energy equals  $E = 0.5mv^2$  such a plot gives the relative energy levels at different values of  $Q$ . In Figure 3.14 the relative energy level is shown of the discussed system with  $f_0 = 100$  Hz. The energy

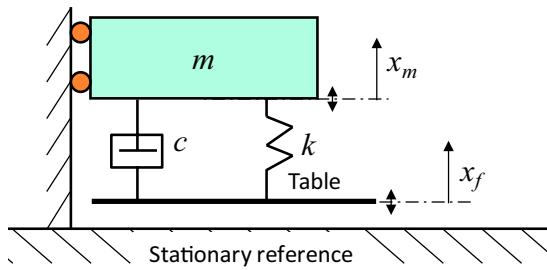


**Figure 3.14:** Relative energy level as function of  $Q$  at a small band around the natural frequency of 100 Hz. The narrow peak at high levels of  $Q$  is especially prominent, when the phenomenon is presented on a linear scale.

at  $Q = 1$  is taken as reference level and it clearly shows, that the maximum energy stored in the system is proportional to the  $Q$  level squared, corresponding with the amplitude of the velocity squared. While it is shown both with logarithmic and linear vertical scales this graphic representation emphasises the very narrow frequency range, also called *bandwidth* at high  $Q$  levels, a property, that is used to its full potential in timing devices like Quartz oscillators, that exhibit  $Q$  factors up to a value of  $10^7$ .

### 3.2.4 Transmissibility of a damped mass-spring system

The term “transmissibility” refers to the capability of a system, to transmit motion from one area to another, both inside a body as well as between connected bodies. It is related to the elastic wave propagation, that was shown in Chapter 2 and is relevant in controlled motion systems, where the path between actuator and sensor always incorporates several flexible elements and bodies with often only limited damping. This effect on control will be further presented in the next section, but first the effect of a motion from a vibrating support, transmitted through the spring and the damper to



**Figure 3.15:** The transmissibility of a dynamic system reflects the sensitivity for external movements of one body, in this case the table, to another body.

a body, is examined. As application example one can think of the inspection microscope on a vibrating table of the first part of this chapter. When the sensitive instrument would be connected to the table by means of compliant springs, the vibrations from the table would be attenuated. This basic *vibration isolation* principle is frequently applied in vibration-sensitive instruments.

To calculate the transfer of movements from the table (\$x\_f\$) to the body with mass (\$m\$), the simple model of Figure 3.15 is used. The distance \$x\_f - x\_m\$ between the two bodies is a constant value in the stationary situation without excitation by floor movements. This constant distance has no influence on the force equations as only the relative displacements will have an impact. For that reason a constant value of zero is chosen for the derivation of the equations of motion.

The total force \$F\_{t,m}\$ acting on the second body by the spring and the damper is defined by the difference in position and velocity between the table and the body and this force induces an acceleration. This force equation is written in the time domain as follows:

$$F_{t,m}(t) = m \frac{d^2 x_m}{dt^2} = c \frac{d(x_f - x_m)}{dt} + k(x_f - x_m) \quad (3.59)$$

After the Laplace transform this force equation equals in the frequency domain:

$$F_{t,m}(s) = ms^2 x_m = cs(x_f - x_m) + k(x_f - x_m) = x_f(cs + k) - x_m(cs + k) \quad (3.60)$$

When shifting the terms the following expression is found:

$$x_m(ms^2 + cs + k) = x_f(cs + k) \quad (3.61)$$

which can be written into the following transfer of  $x_f$  to  $x_m$ :

$$\frac{x_m(s)}{x_f(s)} = \frac{cs + k}{ms^2 + cs + k} = \frac{\frac{cs}{k} + 1}{\frac{m}{k}s^2 + \frac{cs}{k} + 1} \quad (3.62)$$

After introduction of the variables  $\omega_0 = \sqrt{k/m}$  and  $\zeta = c/(2\sqrt{km})$  this transfer function becomes in the frequency domain as function of ( $s$ ) and the radial frequency ( $\omega$ ):

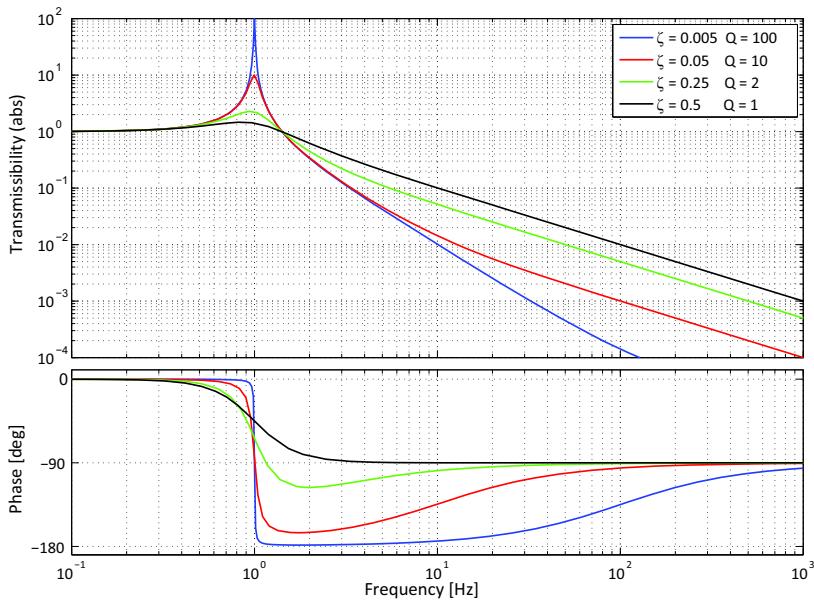
$$\frac{x_m(s)}{x_f(s)} = \frac{2\zeta \frac{s}{\omega_0} + 1}{\frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1} \quad (3.63)$$

$$\frac{x_m}{x_f}(\omega) = \frac{2j\zeta \frac{\omega}{\omega_0} + 1}{-\frac{\omega^2}{\omega_0^2} + 2j\zeta \frac{\omega}{\omega_0} + 1} \quad (3.64)$$

This transmissibility transfer function shows several differences, when compared with the earlier discussed compliance. First of all the equation is dimensionless because it represents the effect of one displacement on another displacement. Furthermore there is no compliance term anymore related to the spring, and at  $\omega \ll \omega_0$  the numerator equals the denominator.

This means, that the transmissibility at very low frequencies always starts at a value of one. The last difference with the compliance transfer function is the presence of an additional differentiating  $s$  term in the numerator that is related to the damper. This term increases the transmissibility proportional with the frequency depending on the damping ratio ( $\zeta$ ). Without this term the transmissibility at higher frequencies would be determined only by the denominator with a  $s^2$  term causing a  $-2$  slope in the magnitude Bode-plot of the transmissibility. For a sensitive instrument that would be placed on the body this would be beneficial, as it reduces the transmission of vibration of higher frequencies to the sensitive instrument. The related  $s$  term of the damper in the numerator however decreases this beneficial effect and this can be understood from the fact that a stronger damper gives a stronger connection between the floor and the body.

These effects are all shown in the Bode-plot of Figure 3.16 where a system is modelled with a natural frequency  $f_0$  of 1 Hz with different damping settings. Even when the damper is not very strong (the red line with  $Q = 10$ ) a rather normal resonance is shown at  $f_0$ , with the  $-2$  slope starting above the natural frequency. For higher frequencies however, the transmissibility



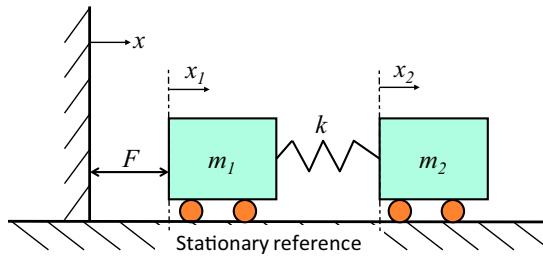
**Figure 3.16:** Bode-plot of the transmissibility of a damped mass-spring system in response to external vibrations from the ground with a natural frequency  $f_0$  of 1 Hz. It clearly shows the negative effect on the reduction of the higher frequencies.

is already much worse with an increase of about a factor ten at hundred Hertz, when compared to the situation with  $Q = 0$  (the blue line). Higher levels of damping ultimately result in a first-order system as can be seen in the phase plot with a  $90^\circ$  phase shift for  $Q = 1$ .

A vibration isolation system with low damping is often used in laboratory equipment, when the natural frequency is sufficiently low to not disturb the measurements. In case of an industrial application, this is often not sufficient. Especially with optical systems that often have natural frequencies in the 100 Hz range additional measures are necessary. In Chapter 9 on wafer scanners, a method will be shown to solve the added transmissibility of the damper by means of active controlled inertia based damping that does not contribute negatively to the transmissibility at higher frequencies.

### 3.3 Multi-body dynamics and eigenmodes

The mechanics in real precision positioning systems are far more complex than the one-dimensional single mass-spring-damper configurations, that were presented in the previous section. When systems must be modelled in multi dimensions, all equations need to be written in a six coordinate direction vector and matrix notation, where also cross couplings between the different directions can be taken into account. These equations often become too complicated for straightforward analytical calculations and in practice they are always done by using computer simulation software like MATLAB and ANSYS. As explained before, this book focuses on the physical understanding of observed phenomena with their basic mathematical modelling, rather than detailed multidimensional calculations. For this reason the use of vectors is restricted to the bare minimum. That being said, even in one direction realistic systems generally consist of a multitude of bodies and springs and it is very important to be able to qualify and quantify the effects of these coupled dynamics from a mechatronic design perspective. First of all, elastically coupled bodies heavily influence the possibility to create a stable controlled motion system, due to their related resonances. Secondly, the measurement position in an active positioning system is hardly ever at the same location as the actuator. For these reasons this chapter will be rounded off by presenting two methods to describe the behaviour of a higher order mechanical system that consists of more than one body connected by springs and dampers. The first method builds on the previous analytical approach with equations of motion, that describe the movement of the bodies. It is shown how these interact with each other. It also becomes clear, that the methodology complicates quickly up to the level, that a real analytical analysis without making errors is not realistic anymore. To cope with this problem, the second method based on the addition of *eigenmodes*, with their corresponding mode-shapes and eigenfrequencies is introduced. This has proven to be a very powerful method to describe the dynamic properties of real complex machines in such a way, that it more easily connects with the design space of mechanical engineers. To illustrate the relation with measurement even further also an example is shown, that works in two dimensions.



**Figure 3.17:** Mass-spring system with two bodies with mass  $m_1$  and  $m_2$  connected by a spring with stiffness  $k$ . A force  $F$  excites the first body and the resulting motion is measured at both bodies.

### 3.3.1 Dynamics of a two body mass-spring system

Mechanical structures can be modelled as a combination of solid bodies connected with springs and dampers. Generally the passive damping of these connections is quite limited and only determined by hysteresis properties of the material or friction in the guiding. For this reason, in the following the behaviour of the dynamic system will be examined without damping in order to avoid long equations. The first example, as shown in Figure 3.17, is a configuration where two bodies are coupled by a spring. The actuation force is applied to the first body with its mass  $m_1$  coupled to the second body with mass  $m_2$  by the spring with stiffness  $k$ . This example is representative for a more realistic dynamic model of the optical pick-up unit of the CD player, where the actuator can be seen as a body with mass  $m_1$  and the lens as a body with mass  $m_2$ .

#### 3.3.1.1 Analytical description

As a first step towards the analytical description of the system dynamics, the balance of forces acting on both bodies is determined along the same reasoning as with the equations of motion of the single body mass-spring system. For the first body the force balance according to the second law of Newton equals in the time domain:

$$m_1 \frac{d^2x_1(t)}{dt^2} = F(t) - k(x_1 - x_2) \quad (3.65)$$

Note, that the force from the spring acts on both bodies in the opposite direction and the external force only works on the first body. For the second body the force balance according to the second law of Newton equals in the

time domain:

$$m_2 \frac{d^2 x_2(t)}{dt^2} = k(x_1 - x_2) \quad (3.66)$$

After Laplace transform these differential equations become in the frequency domain respectively:

$$m_1 s^2 x_1(s) = F - k(x_1 - x_2) \quad (3.67)$$

$$m_2 s^2 x_2(s) = k(x_1 - x_2) \quad (3.68)$$

From these two equations the following transfer functions in the frequency domain between  $x_1$  respectively  $x_2$  and  $F$  can be derived by first using Equation (3.68) to write  $x_2$  as function of  $x_1$  and use that result to solve Equation (3.67):

$$\frac{x_1}{F}(s) = \frac{m_2 s^2 + k}{m_1 m_2 s^4 + k(m_1 + m_2)s^2} \quad (3.69)$$

$$\frac{x_2}{F}(s) = \frac{k}{m_1 m_2 s^4 + k(m_1 + m_2)s^2} \quad (3.70)$$

With these equations some qualitative conclusions can be derived, by looking at different areas of the frequency spectrum. At very small values of  $s$ , the  $s^2$  term in the numerator can be neglected with respect to 1 and in the denominator the  $s^4$  can be neglected with respect to the  $s^2$  term. So for low frequencies both responses become:

$$\frac{x_1}{F}(s) = \frac{x_2}{F}(s) = \frac{1}{s^2(m_1 + m_2)} \quad (3.71)$$

This is a standard mass-line with slope of  $-2$  in the Bode-plot. At high frequencies, so at high values of  $s$  the following approximation is valid:

$$\frac{x_1}{F}(s) = \frac{1}{m_1 s^2} \quad \frac{x_2}{F}(s) = \frac{k}{m_1 m_2 s^4} \quad (3.72)$$

This means, that at these high frequencies the first body will respond according to the mass-line of only its own mass, while the second body will respond with a slope of  $-4$  corresponding with its inability to follow the movement of the first mass. This is called the *decoupling* of the second body.

The transfer functions are written in a polynomial form, that is common in control engineering environments using dedicated mathematical software like MATLAB. To investigate the response at the intermediate frequency range in an analytical way, it is better to write the equations in a different form.

### 3.3.1.2 Multiplicative expression

It is possible to re-arrange Equation (3.69) and (3.70) to obtain a form where a multiplication of factors is used. The multiplicative expression for the transfer function in the frequency domain  $x_1/F(s)$  consists of three terms and equals:

$$\frac{x_1}{F}(s) = \frac{1}{(m_1 + m_2)s^2} (m_2 s^2 + k) \frac{1}{Ms^2 + k} \quad (3.73)$$

With:

$$M = \frac{m_1 m_2}{m_1 + m_2}$$

In this transfer function one can recognise the first term as being the compliance of the free moving bodies together. For small values of  $s$ , corresponding to a low excitation frequency, the two other terms combined are equal to one and the system reacts as one free moving body as concluded in the previous part. The second term  $(m_2 s^2 + k)$  has a constant value equal to the stiffness  $k$  at low frequencies. At high frequencies this factor increases with the square of the frequency, a +2 slope. At a frequency where  $s^2 = (j\omega)^2 = -k/m_2$ , hence  $f = 1/2\pi\sqrt{k/m_2}$  Hz this term becomes zero. As a consequence, the total transfer function will be equal to zero at this frequency. This phenomenon is called an *anti-resonance*, although it is no real resonance, as it does not store energy. The third term is similar to the transfer function of a single mass-spring system.

Apparently there are three frequencies with a dynamic effect. At two frequencies the transfer goes to infinity. These natural frequencies correspond with the poles of the system and are called the *eigenfrequencies*.

$$f_1 = 0 \quad [\text{Hz}], \text{ and } f_2 = \frac{1}{2\pi} \sqrt{\frac{k}{M}} \quad [\text{Hz}] \quad (3.74)$$

The frequency, when the numerator equals zero, determines a zero of this transfer function represented by the 'anti-resonance'.

$$f_a = \frac{1}{2\pi} \sqrt{\frac{k}{m_2}} \quad [\text{Hz}] \quad (3.75)$$

In a similar way as with the first body one can write the transfer function in the frequency domain for the second body:

$$\frac{x_2}{F}(s) = \frac{1}{(m_1 + m_2)s^2} \frac{k}{Ms^2 + k} \quad (3.76)$$

In this equation the same two poles are found, but no zero is found in this transfer function.

### 3.3.1.3 Effect of different mass ratios

In the previous part several dynamic effects were demonstrated occurring with two coupled masses. First the vertical shift of the  $-2$  slope to a higher level from low to high frequencies was shown due to the decoupling of the second mass while the multiplicative notation showed three frequencies with either a resonance or an anti-resonance. To illustrate these findings, the Bode-plots of  $x_1/F$  and  $x_2/F$  will be shown for three different situations:

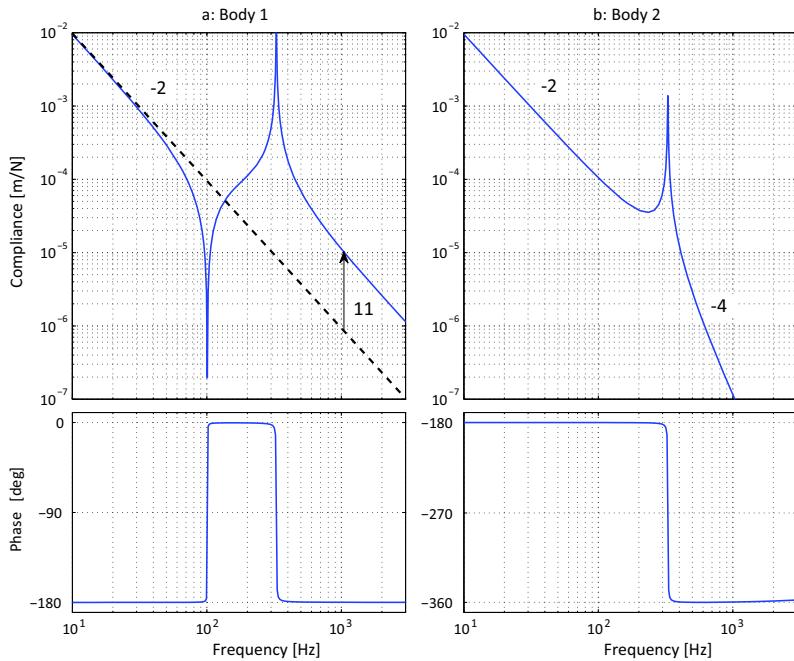
- $m_1 < m_2$ : This is for instance the situation of a body with a large mass ( $m_2$ ), actuated by a less heavy actuator ( $m_1$ ), that is connected by means of a flexible mount.
- $m_1 = m_2$ : This is the case when the mass of an actuator is equal to the mass of the positioned object, for instance when the active mass is optimised relative to the total moving mass of a positioning system.
- $m_1 > m_2$ , This is representative for the situation, where a large mass ( $m_1$ ) is positioned with elastically connected smaller masses ( $m_2$ ), that cause parasitic resonances.

Based on the fact, that  $M$  is always smaller than  $m_2$ , the 'anti-resonance' of the first body will occur at a lower frequency than the resonance at the second eigenfrequency.

Depending on the ratio between  $m_1$  and  $m_2$ , the relation between these frequencies will be different, as shown in Figure 3.18, 3.19 and 3.20.

At  $f = f_a$  the transfer function of  $x_1/F$  shows an 'anti-resonance' according to the zero in the transfer function and all force is directly transferred to the second body. This is a special situation as one would expect a visible resonance at the second body at its "own" natural frequency with the entire spring. This thinking model is however wrong as on the graph of  $x_2/F$  no resonance is shown. This can be explained by the fact, that at this frequency the amplitude  $x_1$  of the first body equals zero and no power is transferred by the force to the second body, which would be necessary to increase the amplitude.

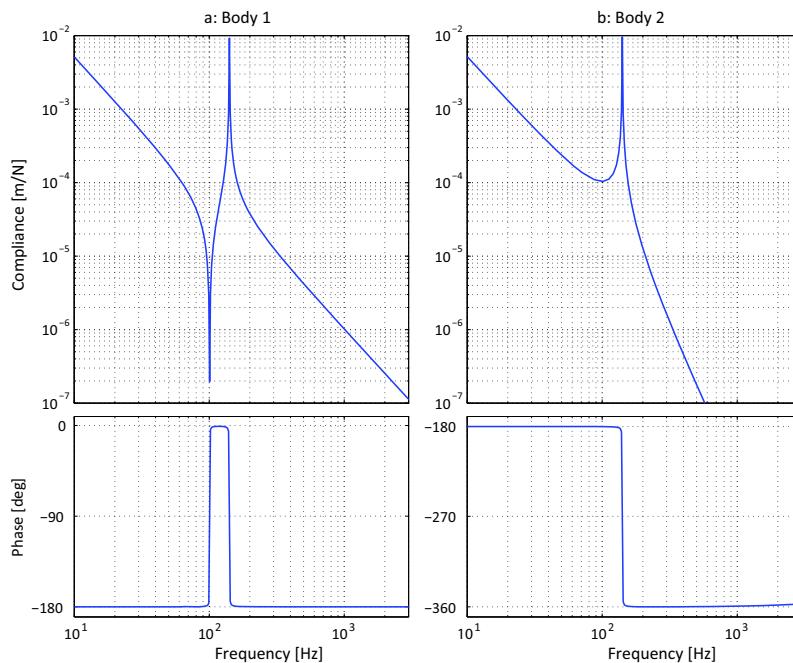
The amplitude of the movement of the second body is such, that the corresponding strain of the spring creates a force that just compensates the external force acting on the first body. This amplitude is equal to the position on the undisturbed line of the Bode-plot and even no reaction is observed in the response of the second body, because at  $f_a$  the amplitude of  $x_2/F$  equals the compliance of spring  $k$  ( $C_s = 1/k$ ). This can be checked in



**Figure 3.18:** Bode-plot of the response of a dual mass-spring system with  $m_1 = 2.5 \cdot 10^{-3}$  kg,  $m_2 = 25 \cdot 10^{-3}$  kg and  $k = 1 \cdot 10^4$  N/m. The values result in a natural frequency of 333 Hz and an ‘anti-resonance’ of 100 Hz for the first body. After the “decoupling” of the second body the initial compliance slope of  $-2$  of the first body continues with the same slope but at higher level, corresponding to the ratio between  $m_1$  and  $m_1+m_2$  ( $=11$ ). At the second body only the resonance is visible and its initial compliance slope of  $-2$  becomes  $-4$  above the resonance, indicating the inability of the second body to follow the movement of the first body.

Figure 3.18.b where the magnitude of the transfer function of the second body at  $f_a = 100$  Hz equals  $10^{-4}$ , which is equal to the compliance of the connecting spring.

At  $f = f_2$  both bodies will resonate while the movement of the first body is  $180^\circ$  out of phase with respect to the movement of the second body as they move in opposite directions. When  $f > f_2$  the slope of the Bode-plot of the first body will continue at  $-2$ , but at a higher level than found by extrapolating the slope at  $f \ll f_a$ . This is caused by the decoupling of the second body, that does no longer join the movement of the first body. As a consequence the slope of the response of the second body becomes twice as steep ( $-2$ ) as the slope of the first body. The compliance of the first body

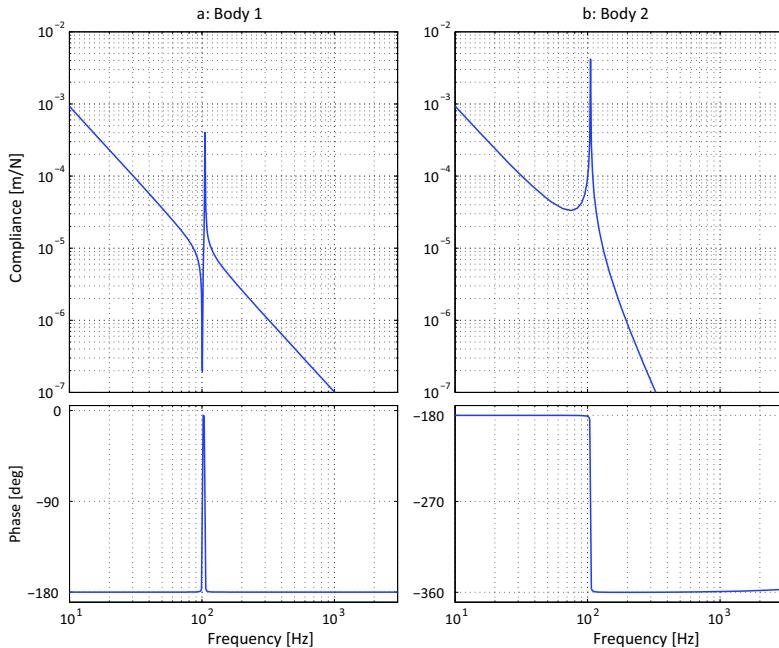


**Figure 3.19:** Bode-plot of the response of a dual mass-spring system with bodies having an equal mass  $m_1 = m_2 = 25 \cdot 10^{-3}$  kg and  $k = 1 \cdot 10^4$  N/m. These values result in a natural frequency of 141 Hz where both bodies move in the opposite direction with equal amplitude. The 'anti-resonance' measured at the first body is not changed compared with Figure 3.18.

is increased proportional to the ratio of the difference in mass, which for a ratio of one to eleven means, that the compliance of the first body becomes a factor eleven larger than was shown in Figure 3.18.a.

For the situation where  $m_1 = m_2$  then  $f_2 = \sqrt{2} \cdot f_a$ . In the Bode-plot of this situation it is shown, that at  $f_2$  both bodies move in counter phase with the same amplitude. As then the middle of the spring does not move, both bodies resonate with only one half of the spring. This double stiffness is the cause of the  $\sqrt{2}$  ratio between  $f_2$  and  $f_a$ .

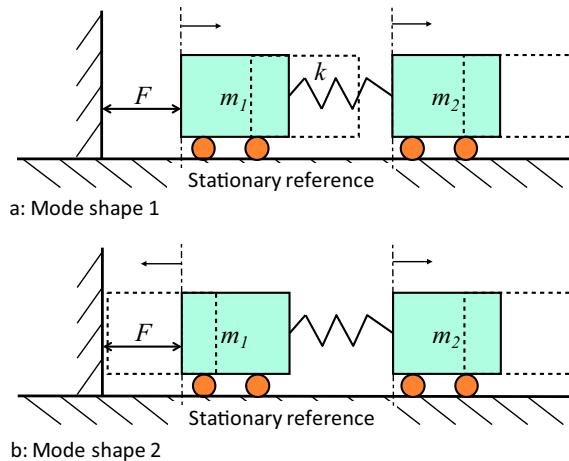
For the situation where  $m_1$  is much larger than  $m_2$ ,  $f_2$  will approximate the value of  $f_a$ . Overall the second body will not have a large influence on the movement of the first body. Only at the natural frequency it will show up as a characteristic combination of a zero and a pole.



**Figure 3.20:** Bode-plot of the response of a dual mass-spring system with the mass of body one being ten times larger than the mass of body two. For this example  $m_1 = 0.25 \text{ kg}$ ,  $m_2 = 25 \cdot 10^{-3} \text{ kg}$  and  $k = 1 \cdot 10^4 \text{ N/m}$ , resulting in a natural frequency of 105 Hz, very close to the 'anti-resonance'. Note, that the overall compliance is decreased due to the increase of the total mass.

### 3.3.2 The additive method with eigenmodes

In the field of Structural Dynamics, the transfer functions from Equation (3.69) and Equation (3.70) are generally written in a different way, derived from the multiplicative expression. The *fourth-order* differential equation that consists of the two coupled second-order differential equations is solved using an eigenvalue decomposition, that leads to two *eigenmodes* with each a natural frequency called the eigenfrequency as presented in Equation 3.74. Associated with eigenmodes also the *mode-shapes* are derived, that visualise the related periodic deformations of these eigenmodes. In Figure 3.21 the example of the previous section is shown, where the first eigenfrequency  $f_1$  is 0 Hz and the corresponding eigenmode has a mode-shape represented by the joint motion of the two bodies. The eigenmode of the other eigenfrequency  $f_2$  has a mode-shape, that is defined by the movement of both bodies in opposite directions, with an amplitude ratio,



**Figure 3.21:** The two eigenmodes with the corresponding mode-shapes of a dual mass-spring system with a connecting spring. The first eigenmode is the linear uniform motion in one direction at the first “eigenfrequency”  $f_1 = 0 \text{ Hz}$ . The second eigenmode is the elastic movement of the two bodies opposite to each other at the second eigenfrequency  $f_2$ . The drawn mode direction is only one of the possible directions as the movement is periodically reciprocating.

that depends on the mass ratio. One can imagine the movements of these eigenmodes as stable oscillations without external forces to keep them going. A constant movement will always continue at the absence of external forces and the same goes for a once excited resonance without damping. It is this straightforward imagination possibility, that makes their use so very valuable in practice. This is even more so because these modes are *independent* as long as the system behaves linear. Under that condition their individual response to a stimulus can be simply added to give to response of the total system. In practical designs this linearisation is often allowed because of the small deformations involved in precision mechatronic equipment.

As a last example of the analytic equations of motion, the related two transfer functions are derived to illustrate this superposition principle. Starting with a slightly different notation of Equation 3.73 the following equation is obtained for the position of the first body in the frequency domain:

$$\frac{x_1}{F}(s) = \frac{1}{(m_1 + m_2)} \frac{1}{s^2} \frac{m_2 s^2 + k}{M s^2 + k} \quad (3.77)$$

To convert this into two terms with  $1/s^2$ , that can be added, as a first step  $M s^2$

will both be added and subtracted from the numerator of the last equation:

$$\begin{aligned}\frac{x_1}{F}(s) &= \frac{1}{(m_1+m_2)} \frac{1}{s^2} \frac{(Ms^2+k)+(m_2-M)s^2}{Ms^2+k} \\ &= \frac{1}{(m_1+m_2)} \left( \frac{1}{s^2} + \frac{m_2-M}{Ms^2+k} \right) \\ &= \frac{1}{(m_1+m_2)} \left( \frac{1}{s^2} + \frac{\frac{m_2}{M}-1}{s^2+\frac{k}{M}} \right)\end{aligned}\quad (3.78)$$

With  $M = m_1 m_2 / (m_1 + m_2)$  and the natural angular frequency of the second mode-shape  $\omega_2 = \sqrt{k/M}$ , the total transfer function becomes written as the addition of the transfer functions of two separate mass-spring systems:

$$\frac{x_1}{F} = \frac{C_1}{s^2} + \frac{C_2}{s^2 + \omega_2^2} \quad (3.79)$$

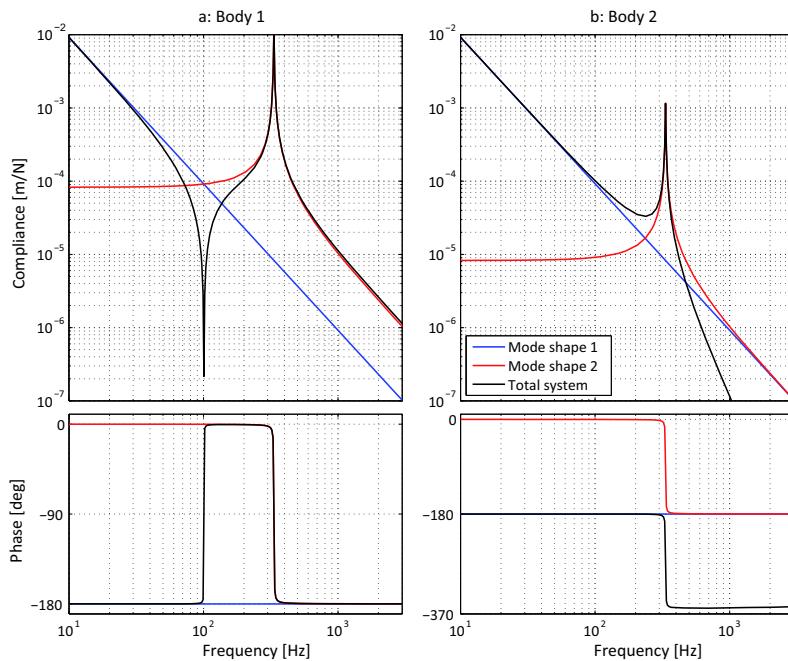
With:

$$C_1 = \frac{1}{m_1 + m_2} \quad \text{and} \quad C_2 = \frac{1}{m_1} - \frac{1}{m_1 + m_2} = \frac{1}{m_1 + m_2} \frac{m_2}{m_1} \quad (3.80)$$

The same “transformation” steps, based on Equation 3.76, result in the analytical transfer functions for the second body in the frequency domain.

$$\begin{aligned}\frac{x_2}{F}(s) &= \frac{1}{(m_1+m_2)} \frac{1}{s^2} \frac{k}{Ms^2+k} \\ &= \frac{1}{(m_1+m_2)} \frac{1}{s^2} \frac{\left(s^2 + \frac{k}{M}\right) - s^2}{s^2 + \frac{k}{M}} \\ &= \frac{C_1}{s^2} - \frac{C_1}{s^2 + \omega_2^2}\end{aligned}\quad (3.81)$$

With this transformation Equation (3.79) and Equation (3.81) appear to consist of two simple dynamic transfer functions, that are superimposed. The first term is the typical compliance of a single mass and corresponds with the first mode-shape. The second term corresponds with a second-order mass spring system and becomes similar to Equation (3.32) by dividing both numerator and denominator by  $\omega_2^2$ . When looking at the difference in both equations, it can be concluded, that for the first body, where the external force is acting, the two transfer functions are **added** with **different** compliance factors. On the other hand for the second body the second mode is **subtracted** from the first mode with **equal** compliance factors. This difference exists, because in the second mode both masses move in opposite



**Figure 3.22:** Bode-plot of the response of the same dual mass-spring system of Figure 3.18 with  $m_1 = 2.5 \cdot 10^{-3}$  kg,  $m_2 = 25 \cdot 10^{-3}$  kg and  $k = 1 \cdot 10^4$  N/m, resulting from the combination of the two mode-shape responses. The response of the first body (a:) is the result of the addition of both mode-shape responses, while the response of the second body is the result of the subtraction of the second mode-shape from the first mode-shape. The difference between the responses of both bodies in the spring-line of the second mode-shape is the direct consequence of the difference in mass. The largest mass  $m_2$  will get the smallest part of the spring corresponding with a lower compliance. Also the 'anti-resonance' appears to be no resonance at all, as it is the result of the combination of two equal amplitudes with  $180^\circ$  phase difference.

directions with an amplitude ratio, that is inversely proportional to the ratio between the masses of the two bodies.

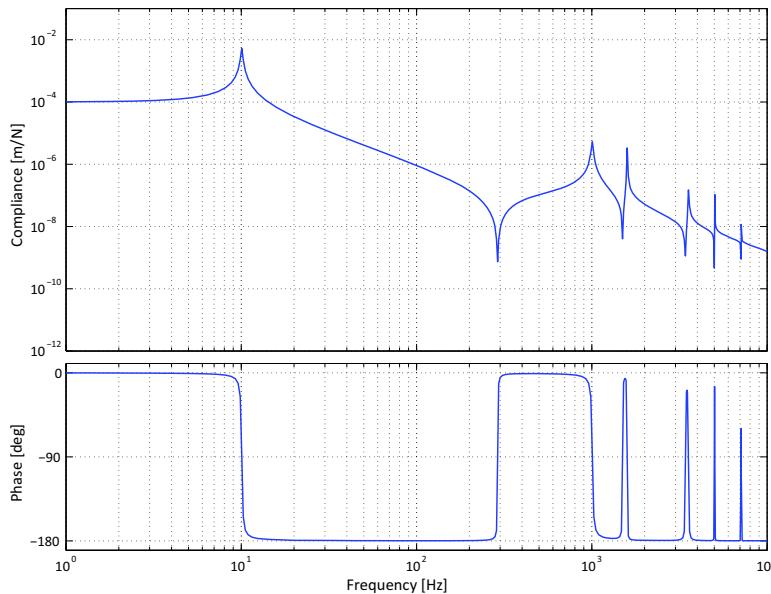
As a result of this different formulation of the transfer functions, the Bode-plot of a complex system can be relatively simply be derived by combining the Bode-plots of the separate eigenmodes. The result is shown in Figure 3.22, that is identical to Figure 3.18 in respect to the response of the total system. For the response of the first body, at the frequency, where both contributions have an equal amplitude, they have an opposite phase,

and thus the two contributions will add to zero. This makes clear, that the term 'anti-resonance' is badly chosen as it is only the result of two equal signals in counter phase. Similarly the transfer function  $x_2/F$  of the second body can be determined out of the two separate transfer functions. In the second mode-shape the second body moves in the opposite direction from the first body, which means, that the response of the second mode-shape has to be subtracted from the response of the first mode-shape to achieve the total response. Due to this phase inversion the anti-resonance does not occur. At higher frequencies beyond  $f_2$  the phase of the second eigenmode is  $180^\circ$  delayed. Because then both components have approximately the same magnitude and an opposing phase, the combined magnitude goes to zero, leading to the  $-4$  slope for high frequencies with the corresponding  $-360^\circ$  phase relation.

### 3.3.2.1 Multiple eigenmodes and modal analysis

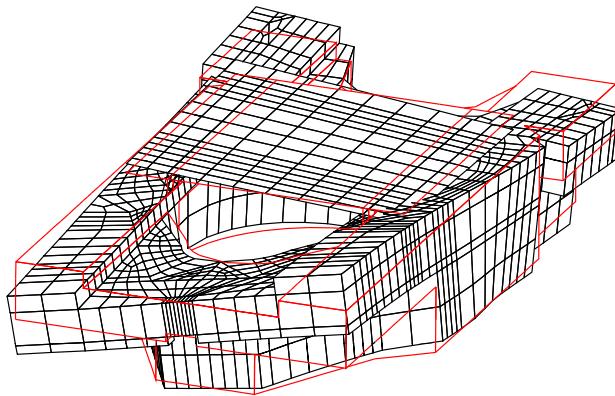
The example of the previous section consisted only of a simple two body mass-spring system, without a connection to the stationary world. Generally the dynamics of a positioning system are more complex. When for instance a spring is added in Figure 3.17, that connects the first body to the stationary reference, the transfer function of the first mode-shape becomes  $C_1/(s^2 + \omega_1^2)$ , similar to the second mode. When measuring the real dynamic transfer function of a mechanical system, one could obtain a Bode-plot, like shown in Figure 3.23. The first mode corresponds with the mass of 2.5 kg of this example system, connected by a spring to the stationary world with a stiffness  $k = 10^4$ . At higher frequencies several other resonances are visible, each corresponding to a different eigenmode, caused by several parts connected to the main body by supports with a certain stiffness. All of these eigenmodes add another two orders to the analytical differential equation and this complicates the analysis considerably. Even without a thorough higher-order analysis it is often possible to estimate the cause of these eigenmodes, given the mass of the different parts in the structure. With this information the mechanical designer can adapt the dynamic behaviour to the required properties.

Unfortunately, frequently the resonances are due to three dimensional deformations of the body itself, like shown for instance in Figure 3.24. In such cases Finite Element Modelling of the dynamics, with software like ANSYS, is necessary to calculate the eigenfrequencies and mode-shapes of each eigenmode. Practical verification is done with a method called *modal analysis*. In this method the response to a dynamic stimulus is measured



**Figure 3.23:** Bode-plot of a realistic dynamic transfer function of a mechanical system. The total mass of this example is 2.5 kg and the main body is connected to the stationary world by a spring with a stiffness  $k$  of  $10^4 \text{ N/m}$ . The natural frequencies in the graph each correspond with different eigenmodes, that are determined by parts that are attached to the main body by means of a mount with a certain stiffness. The measurement and the actuation are on the same location, resulting in a phase between  $0^\circ$  and  $-180^\circ$ . Measuring the position at one of the decoupled masses would cause much more phase delay.

at different locations on the subject by means of accelerometers or other sensitive sensors for measuring short range displacements. The stimulus often consists of a force impulse, delivered by a calibrated “hammer”. An impulse stimulus contains a wide range of frequencies as was presented in Chapter 2. The different responses are compared with each other and with the frequency spectrum of the stimulus by means of correlation algorithms, that derive the related mode-shapes and amplitudes. Presently enhanced correlation methods even enable this modal analysis without a separate stimulus, by just using the random external vibrations, that are often present in actual systems.



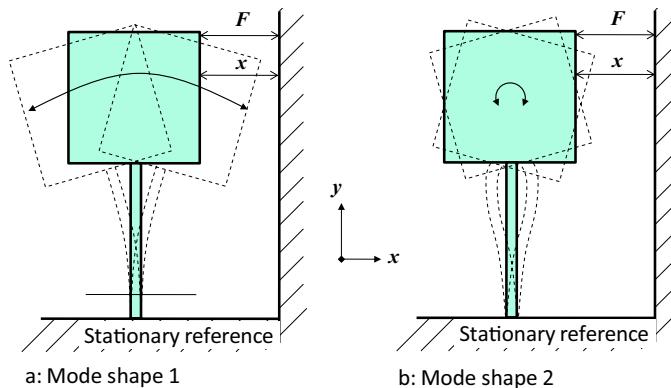
**Figure 3.24:** One of several mode-shapes of a solid body as calculated with Finite Element Modelling software. The red lines indicate the undeformed shape. Excitation of this torsional eigenmode is often extremely difficult to avoid by other means than a careful lay out of the actuator and sensor locations.

### 3.3.2.2 Location of actuators and sensors

In order to link this chapter with the following chapter on motion control, the negative effect of eigenmodes on the dynamic performance of a controlled motion system will be elaborated a little more from the aspect of phase in relation to the position of actuators and sensors. These positions have their influence on the *observability* and *controllability* of an active controlled system which refers to the possibility to measure and control the eigenmodes of a dynamic system.

In the previous sections on the two body mass-spring system it was shown, that measuring the position at the second body results in a maximum phase shift of  $-360^\circ$ . This is far above the maximum of  $-180^\circ$ , that is required for stability in a feedback loop, as will be shown in the next chapter. Also Figure 3.23 might look all right from a phase perspective but as soon as the measurement system is connected to one of the decoupling masses instead of directly at the actuator, the resulting phase delay would be significant. This phase problem by eigenmodes is also observed in Figure 3.24, as the shown torsional mode represents a movement, that is different for all parts of the body both in phase and amplitude.

With the more simplified example of Figure 3.25 it will be illustrated, how these eigenmodes and their corresponding frequency response measurement are impacted by the location of the actuators and sensors. It emphasises the



**Figure 3.25:** The mode-shapes of the two main, in plane, eigenmodes of a body, connected to the stationary reference by means of a leaf-spring. The first mode-shape consists of a swinging movement around a pole close to the stationary reference and the second mode-shape is a reciprocating rotation around the centre of mass. Depending on the location of actuation ( $F$ ) and sensing ( $x$ ), the observed dynamic behaviour of the system is different, observed primarily with the second mode-shape.

importance of the right choice of this location in an actual design. This two dimensional model consists of a rigid body connected to the stationary reference by means of a mass-less ideal leaf-spring with only two mode-shapes of interest, a “swinging” movement around a pole close to the connection of the leaf-spring with the stationary world and a rotation around the centre of mass of the body. The method to combine the response of eigenmodes of a single body is identical to the method of the combined eigenmodes of a multiple body system, because in both cases the mode-shapes of the different eigenmodes are independent. This means, that they can be linearly combined to give the resulting overall behaviour, as long as the coordinate system of all mode-shapes is identical. This does not necessarily mean, that the results are identical and it will be shown, that especially with different locations of actuator and sensor the response becomes different.

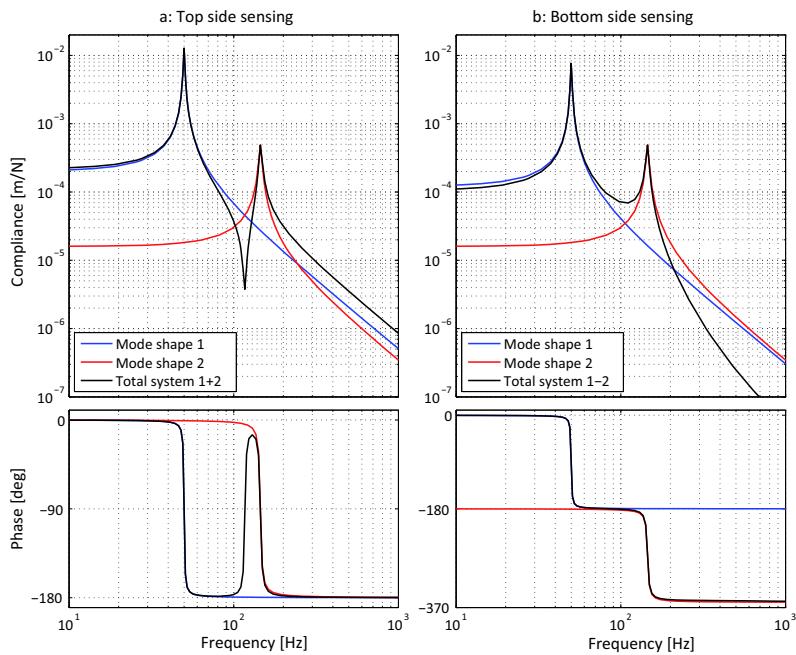
A periodic force stimulus  $F$  with frequency  $f$  is exerted on the top side of the body. To illustrate the difference in perceived behaviour, the measurement of the movement is done at different positions between the bottom side and the topside of the body.

For the calculation of the eigenfrequencies and the Bode-plot, some values are chosen as example for the different parameters. The mass ( $m$ ) of the body equals 0.05 kg and the bending stiffness ( $k$ ) of the leaf-spring for the

first mode-shape is  $5 \cdot 10^3$  N/m. These values result in an eigenfrequency of 50 Hz of the first eigenmode. The eigenfrequency of the second eigenmode is determined by the moment of inertia around the centre of mass in combination with the rotation stiffness of the leaf-spring. As only the resulting effect in the  $x$  direction is important, these values can be transformed into an equivalent mass and stiffness, located at the point of contact of the excitation force. When the equivalent mass is calculated at the corner of the square, a value is found of approximately 0.033 kg. It is not without logic, that the rotational stiffness of the second mode-shape is significantly higher than the bending stiffness of the leaf-spring for the first mode-shape. For this reason the equivalent stiffness at the point of exertion of the force is chosen to be  $3 \cdot 10^4$  N/m in order to result in a eigenfrequency of approximately 150 Hz. When the position measurement is done at the point of exertion of the force, the observed behaviour is shown in the Bode-plot of Figure 3.26.a. The responses of both eigenmodes are according to the response of a standard mass-spring system and the combination results in a comparable response as with a coupled dual mass-spring system, as shown before in Figure 3.22. The only difference with the previous example is the eigenfrequency of the first eigenmode, that is 50 Hz instead of zero. The same 'anti-resonance' is observed and it is shown, that a single flexible body can behave dynamically comparable as a multi-body object connected with springs.

A totally different behaviour is however observed when the measurement takes place at another location. When the motion is measured at the lower side of the body, the resulting response is shown in Figure 3.26.b. For the first eigenmode the magnitude of the response is reduced, when compared with the measurement on top of the body, because the measurement takes place closer to the mechanical pole of the swinging movement. The largest and most relevant difference is however found in the second eigenmode, as the magnitude is equal, but the phase is 180° shifted in respect to the movement at the top. For this reason the combination of these responses by addition becomes a subtraction. As a consequence at the frequency, where both mode-shapes have an equal amplitude, the phase of both signals is -180° and gives a higher value instead of the anti-resonance. At higher frequencies above the second eigenfrequency the slope of this combined response becomes -4 with a corresponding -360° phase. While this would be comparable with the behaviour of the second body of the coupled dual body mass-spring system, there is a difference. Instead of only the second eigenfrequency, both eigenfrequencies are visible, which is due to the fact, that the phenomenon is acting on one single body.

The conclusion is, that depending on the location of the position sensor,



**Figure 3.26:** Bode-plot of the response of the two mode-shapes of Figure 3.25. The left graph (a:) shows the measured behaviour when sensing at the same location as the force, while the right graph (b:) shows the measured response when sensing at the bottom of the body. In that situation the 'anti-resonance' is disappeared but both resonances remain present.

the observability of the mode-shape is changed. This is especially the case, when the measurement sensor is located in the middle, where the movement in the  $x$  direction due to the second mode-shape becomes zero. At this point this eigenmode will not be observed anymore by the sensor. When this configuration is applied in a closed loop feedback system, the dynamic effects of the second eigenmode will not interfere with stability, which is a positive property. Unfortunately this does not mean, that this eigenmode is not excited, as unknown external forces, due to noise from the actuator or vibrations from outside, could contain frequencies around the eigenfrequency of this not observed mode-shape.

For these reasons, in a real design it is always necessary to prevent these unknown forces as much as possible, for instance by means of vibration isolation. It is also necessary to make sure that the actuator is not capable of exciting an un-observed mode-shape by following the same reasoning as

above for the position of the actuator. As long as in the example this position is in line with the centre of mass, the second eigenmode will not be excited by the actuator. This optimal position from the point of view of avoiding resonances in eigenmodes comes at a price in a reduced controllability of the system. Even if the controller would have information about this eigenmode it will not be able to control it because it is missing an actuator at the right location. This can be solved with a separate actuator, only intended to control this eigenmode, but more often very strict measures are taken to prevent any other source of interference. This is one of the many important attention points, when designing complex mechatronic systems with extreme precision.

### 3.3.2.3 Summary

In this chapter some important lessons can be learnt, that are summarised as follows:

- Stiffness, whether it is created mechanically or by means of a control system, is determinative for precision.
- Every mechanical structure can be modelled as a combination of bodies, springs and dampers, either as separate bodies or as finite elements within a body.
- All body-spring combinations determine a mass-spring system with damping with its related natural frequency and phase delay.
- Phase is a prominent factor regarding the possibility to control the system.
- Compliance and transmissibility are similar but not equal.
- The quality factor  $Q$  and damping ratio  $\zeta$  are interchangeable. Each have their practical value.
- A damper is necessary to control the resonance at the natural frequency but it affects the response for transmissibility at the higher frequencies.
- A complex system can be modelled as a simple combination of its eigenmodes, each with its own mode-shape, eigenfrequency, damping and phase behaviour.
- The position of the actuator and the sensor determine the observability and controllability of the different eigenmodes.

- A precision design requires a careful lay-out of all elements, considering the eigenmodes.

Finally it can be concluded, that these insights help in designing actively controlled dynamic systems with optimally located actuators and sensors, that reduce the sensitivity for modal dynamic problems. With this conclusion this chapter closes to continue with the active elements of a mechatronic system.

# **Chapter 4**

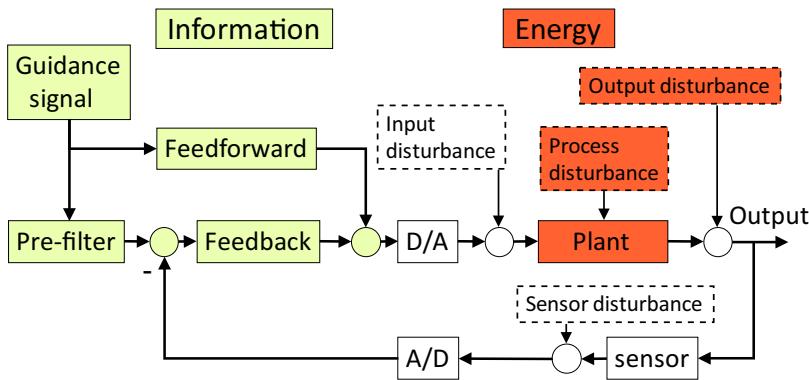
# **Motion Control**

## **Introduction**

As was presented in the previous chapters, most mechatronic systems are actively controlled motion systems, which implies that these systems are of a dynamic nature. If the mechatronic system to be controlled does not show any dynamics within the required positioning bandwidth, the entire control problem becomes quasi static and is therefore trivial.

In most motion control systems this is not the case and compensation of the system dynamics via control is required to achieve the specified performance in terms of precision, accuracy and frequency response. This chapter discusses the various approaches to guide and actively control motion systems. As the name already indicates, motion control is all about the control of a machine to follow a pre-defined trajectory in space and time, with various applications. Examples are precision position control with rejection of disturbances due to vibrations from the environment or imperfections of the mechanical system as well as path planning and velocity control for scanning applications.

This chapter is divided in three parts. First the control loop will be examined to give a global overview. The second part explains in more detail the more "classical" control design of feedforward and PID-feedback control with a strong emphasis on the frequency response and the transfer functions. The last part will give a short overview of the state-space representation that is used in modern, model based control designs and allows to model the system in the time domain..



**Figure 4.1:** Block diagram of a motion control system, including feedforward and feedback control. The plant consists of the power amplifier, actuator and the mechanical dynamics. The diagram clearly shows the different places where interfering disturbances impact the system.

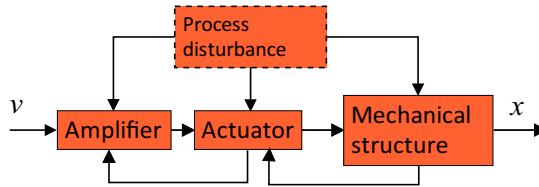
## 4.1 A walk around the control loop

Figure 4.1 shows a basic control loop of a positioning system where the control part consists of a feedforward path and a feedback path. Control systems could in principle consist only of a feedforward or feedback path, but generally a combination is applied as each principle has its own advantages and disadvantages, that are separately discussed in the following sections. For simplicity this chapter is limited to linear dynamics which fortunately holds for most practical precision motion control systems. In most cases also a non-linear motion system can be linearised around a chosen operating point.

The *plant* is a control engineering term for the uncontrolled physical system. In a positioning system this consists of the power amplifier, actuator, and the mechanical structure, as shown in Figure 4.2.

Each of these elements has its own inherent dynamic properties. They interact in both directions in such a way that each element not only determines the input of the next element but also influences the previous element by its dynamic load. Often the plant is called just simply the “system”, which can be confusing when it is not specified, as it can mean the controlled or uncontrolled system, the control system or whether it includes the sensor or not.

Also in this book the term ”system“ is more generally used and hence not precisely defined. It can for instance be the mechanical mass-spring system



**Figure 4.2:** In a mechatronic system the plant consist of the amplifier, the actuator and the mechanical structure. Sometimes also the sensor is included with the related electronics. The elements that transfer energy interact with each other in two directions. For instance the amplifier determines the input of the actuator, while the actuator determines a dynamic “load”, that influences the behaviour of the amplifier. The same goes for the actuator and the mechanical structure.

from the previous chapter, the control system or even the full controlled mechatronic system including the sensor. Its meaning will however be clear from the context.

As shown in Figure 4.1, the functional blocks that determine the plant with the related disturbances are represented in red and the functional blocks forming the control-system are represented in yellow.

It is necessary to clearly distinguish the domain of each functional block, because everything that is located at the left side from the D/A and A/D-converters, including the converters themselves, is only involved with information exchange. This means that these blocks can be interconnected in serial or parallel configurations, without changing the properties of the individual block. On the other hand however, the blocks that are shown in red represent physical systems or sub-systems where energy exchange is involved. This implies that changes in one sub-system (or block) may also change the dynamics of the blocks that it is interacting with. A simple example is a motor that behaves differently whether a mechanical load is connected to it or not. These aspects of system thinking have to be considered carefully when interconnecting mechatronic sub-systems, including the control system.

For analysis and discussion of the control part, first a short section on *poles* and *zeros* will determine the red thread through the chapter, then *feedforward* and *feedback* control will be presented separately.

### 4.1.1 Poles and zeros

In Section 3.2.3 of the previous chapter the pole of a transfer function was introduced for those values of the Laplace parameter  $s = \sigma \pm j\omega$  where the denominator of the transfer function becomes equal to zero. It was used to explain the relation between the location of the poles in the Laplace plane and the damping ratio  $\zeta$  of the mass-spring system. It was shown that, when the poles are located in the left-half of the Laplace plane, the ratio between the real part and the imaginary part will determine the amount of damping in the mass-spring system.

Many passive mechanical systems, like the damped mass-spring systems from the previous chapter, are *stable* in the uncontrolled, *open-loop* situation. In dynamic systems *stability* is defined by the ability of the system to be insensitive to external stimuli or to return to a defined equilibrium situation after the application of a stimulus.

#### 4.1.1.1 Controlling unstable mechanical systems

There are also examples of passive mechanical systems that are not stable, like an inverted pendulum or a piece of iron between two permanent magnets. These systems can not be stably positioned at any location between two (or more) extreme positions as they show a *negative stiffness*. To illustrate the effect in the frequency domain the transfer function can be derived from the general compliance transfer function of a standard mass spring system. The compliance without damping was shown to be equal to:

$$C_t(s) = \frac{x}{F} = \frac{1}{ms^2 + k} \quad (4.1)$$

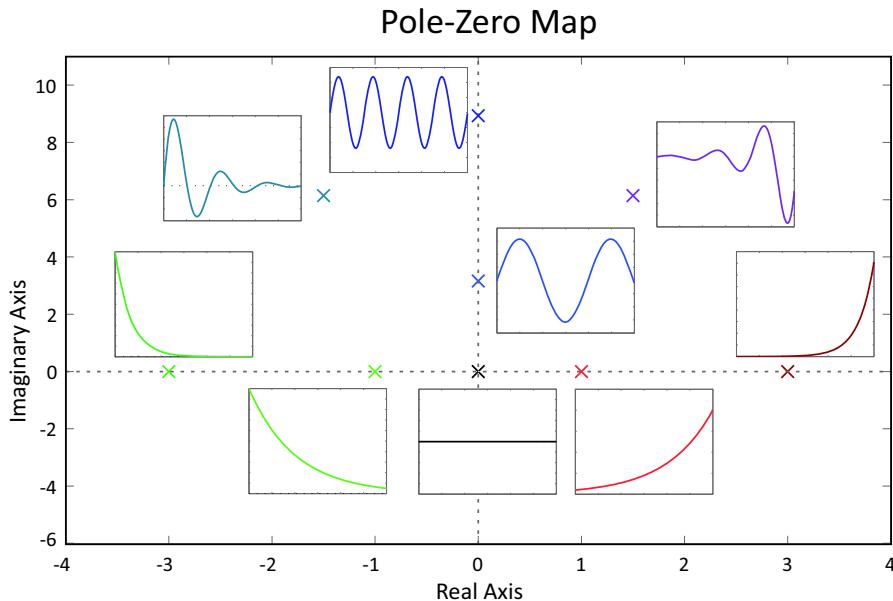
A spring with a negative stiffness implies that  $k$  has a negative value and as a consequence the two poles of this transfer function are both real with a value of  $\pm\sqrt{|k/m|}$ . One of these values is positive which means that the location of this pole is on the real axis in the right-half of the Laplace plane, indicating an unstable pole. The related instability is manifested by a increasingly accelerated movement away from the zero position as is shown in Figure 4.3. With feedback control such an unstable system can be stabilised by shifting the positive pole from the right to the left-half of the Laplace plane. This effect can be explained with the example of Chapter 3 where proportional negative feedback is shown to create a spring with a positive stiffness for the positioning of an optical pick-up unit of a CD player. When this principle of negative feedback is applied on a mechanical system with a negative stiffness

like the *magnetic bearing* that will be presented later in this chapter, this positive stiffness can compensate the negative stiffness when the total loop gain is sufficiently high. When the positive stiffness by the feedback is even higher than the negative stiffness of the mechanical system, the resulting positive stiffness creates a normal mass-spring system with two conjugate complex poles on the imaginary axis and this transformation implies that the originally positive pole is shifted towards the left-half of the Laplace plane by the negative feedback. The addition of damping by differentiating the feedback signal, as will be explained later, will move the poles even further to the left and such a magnetic bearing will behave just like any other well-controlled positioning system.

#### 4.1.1.2 Creating instability by active control

With active dynamic systems the presence of elements that add energy to the plant like amplifiers and actuators can also induce instability in a normally stable passive system, as can also be explained with a simple example. In the previous chapter it was demonstrated that a negative proportional feedback loop creates a system with the same behaviour as would occur with a mechanical spring. This means that by only reversing the sign ( $= 180^\circ$  phase) of this feedback loop, a negative stiffness would be created with the inherent unstable pole in the right-half of the Laplace plane.

In dynamic systems the phase in the loop is a function of the frequency and in mechanical systems phase lag is present due to the influence of masses and damping and higher order eigenmodes. Ultimately these effects result in a large enough phase lag to turn the system into an unstable situation even with negative feedback. For this reason the analysis of the poles is of prime importance when dealing with active control of dynamic systems. Figure 4.3 gives an overview of the impulse response of a dynamic system for different locations of the poles in the Laplace plane. It clearly underlines the observed phenomena. The location on the real axis corresponds with non-oscillating gradually changing responses, either to a stable position for poles in the left-half-plane or to infinity for poles in the right-half-plane. The imaginary part of the pole determines the oscillatory behaviour in relation to the real part. The magnitude of the imaginary part ( $j\omega$ ) determines the oscillation frequency and the ratio between the imaginary part and the real part determines the damping ratio.

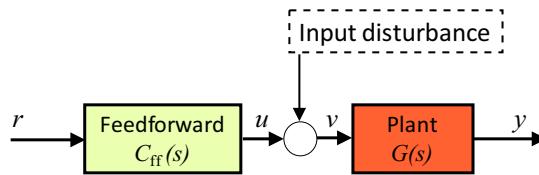


**Figure 4.3:** The impulse response of a dynamic system as function of the pole locations in the Laplace plane. In the left-half-plane the response returns asymptotically to a zero deviation value while in the right-half-plane the response grows exponentially to infinity. For simplicity only the positive imaginary axis is shown but any pole with an imaginary part also has a conjugate complex counterpart mirrored around the real axis.

#### 4.1.1.3 The zeros

The counterpart of the poles are the zeros. They are not detrimental to stability as they correspond with a value of  $s$  where the numerator of the transfer function becomes equal to zero. As an example, the anti-resonance of the coupled mass-spring system is a pair of conjugate complex zeros of the transfer function, as shown in Equation (3.73) in the previous chapter. The drawback of a zero is the lack of reaction to a stimulus at that frequency. It can, however, be beneficially used in feedforward control because zeros can compensate poles as can be seen at the following generalised transfer function written in pole-zero notation where  $p$  represent the poles and  $z$  represent the zeros:

$$G(s) = \frac{(s - z_1)(s - z_2)(s - z_3)(s - z_4)(s - z_5)\dots(s - z_m)}{(s - p_1)(s - p_2)(s - p_3)(s - p_4)(s - p_5)\dots(s - p_n)} \quad (4.2)$$



**Figure 4.4:** Block diagram of a feedforward-controlled motion system with one input and output (SISO). The transfer functions  $C_{\text{ff}}(s)$  of the feedforward controller and  $G(s)$  of the plant are frequency dependent, which is denoted by the Laplace variable  $s$ .

where  $m$  and  $n$  are integers with  $m \leq n$ .

An equal pole and zero cancel each other out and this can be used to improve the dynamic behaviour of the system. Furthermore a single real zero in the left-half plane gives a phase lead of  $90^\circ$  and a  $+1$  slope in the Bode-plot, corresponding to a differentiating behaviour.

In the following sections the existence and locations of poles and zeros will be frequently addressed.

## 4.1.2 Properties of feedforward control

Figure 4.4 shows the typical basic configuration for feedforward control, which sometimes is also called *open-loop control*. This example has only one input and output variable which is in control terms a *Single Input Single Output* (SISO) system.

It is important to note that in mechanical engineering a SISO positioning system is also called a single degree of freedom system, relating to a specific direction in an orthogonal three dimensional spatial coordinate system. Unfortunately the term “degree of freedom” can lead to confusion in position control of mechanical systems, because in control terms, feedforward and feedback are considered as two degrees of freedom even though it relates to one direction. For that reason the term is avoided further in this chapter. The reference or guidance signal<sup>1</sup>  $r$  is applied to the controller that has a frequency dependent transfer function  $C_{\text{ff}}(s)$ . The output  $u$  of the controller is connected to the input of the motion system that has a transfer function

<sup>1</sup> According to the domain notation that was defined in Chapter 2 the  $(t)$ ,  $(s)$ ,  $(\omega)$  and  $(f)$  terms, that define whether the equation is in the time or the frequency domain, are only mentioned once before the equal sign. One exception to this rule is applied: When transfer functions are included without expanding them, these functions will also show an indicating term in order to distinguish them from time invariable functions. The terms are never shown with variables because these are variable by nature according to the domain that the function describes.

$G(s)$  giving the output  $y$ , which is a position. In this configuration the feedforward controller acts as a filter that modifies the reference signal in such a way, that the motion of the controlled mechatronic system follows the reference signal.

If one would like to achieve perfect control, which means that there is no difference between the reference position and the actual position of the system, the combined transfer function  $G_{t,ff}(s)$  from  $r$  to  $y$  has to be equal to one, hence show identity:

$$G_{t,ff}(s) = \frac{y}{r} = C_{ff}(s)G(s) = 1 \quad (4.3)$$

In that case the feedforward controller has to be the exact inverse of the plant

$$C_{ff}(s) = G(s)^{-1}. \quad (4.4)$$

If no dynamics are involved, the feedforward controller eventually would only represent a gain that scales the reference signal. In reality positioning systems include dynamics with a frequency dependent transfer function. In that case also the dynamics of the positioning system have to be inverted, which results in *pole-zero cancellation* between the controller poles and system zeros as well as controller zeros and system poles.

Feedforward control is a very useful and often necessary first step in the control of a complex dynamic motion system as it provides the following advantages:

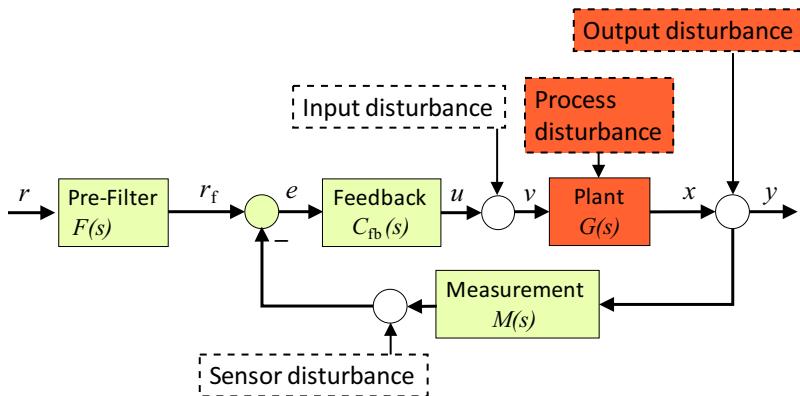
- **No sensor required:** No sensor information is fed back to the system which means that a sensor can be left out, thus reducing the cost of pure feedforward controlled systems.
- **Predictable movement:** If the reference signal (trajectory) is known in advance, the phase lag and time delay in the system can be predicted and therefore compensated.
- **No introduction of instability:** The poles of the controlled system are not changed by feedforward control. Therefore no instability can be introduced under the trivial condition that the feedforward controller itself is stable.
- **No feeding back of sensor noise:** In precision motion systems, positioning noise is a critical point that always has to be considered in the control design. The lack of a sensor avoids insertion of the measurement noise in the system.

The feedforward problem can be more complicated as not always all plant dynamics can easily be inverted. For example a system with a low pass characteristic, a property of all positioning systems that are limited by the mass-line, would require a feedforward controller with high-pass characteristics, having a very high gain at high frequencies. From this reasoning it is obvious that feedforward control of motion systems is applicable for a certain frequency range only, depending on the system dynamics. A second problem is that an unstable system cannot be stabilised with open-loop control by cancelling the unstable pole of the system. Further also unstable transfer zeros, non-minimum phase zeros with a not unique phase to magnitude relationship, must not be inverted as they would lead to an unstable pole in the feedforward controller.

This all is included in the following drawbacks and limitations of feedforward control:

- **Limitation to inverted low pass characteristic** It is not possible nor wise to create a controller with a very high gain at very high frequencies.
- **The plant has to be stable:** Unstable systems cannot be controlled with pure feedforward control. The smallest disturbance or noise, which is always present, would cause the system to become unstable even if an unstable pole would be cancelled by a corresponding zero.
- **No compensation of model uncertainties:** Variations in the system dynamics, such as shifting of the resonance frequency or variation of the damping, are not monitored and therefore not accounted for in feedforward control.
- **Can only compensate for known disturbances:** Disturbances of the motion system can only be compensated, if they can be measured. Even with additional measurement these disturbances can only be compensated to a certain limit, because for perfection the influence of the control should be at least as immediate at the systems output as the disturbance. This implies an infinitely fast reaction which is practically impossible.

In motion control of mechatronic systems feedforward control is a very important component mainly because it is faster than pure feedback control.



**Figure 4.5:** Block diagram of a SISO feedback controlled motion system. Each functional part has its own, mostly frequency dependent transfer function.

### 4.1.3 Properties of feedback control

In feedback control the actual status of the motion system is monitored by a sensor and the controller is generating a control action based on the difference between the desired motion (reference signal) and the actual system status (sensor signal).

Figure 4.5 shows the block diagram of a standard SISO feedback loop. The output<sup>2</sup>  $y$  is measured and compared with (subtracted from)  $r_f$  which is the reference  $r$  after filtering. The result of this comparison is used as input for the feedback controller.

Because the sensor signal is fed back in a closed-loop to the input of the system, feedback control is also called *closed-loop* control.

The transfer function of a feedback loop is derived from the following equations in the frequency domain:

$$e = r_f - M(s)y, \quad y = G(s)C_{fb}(s)e, \quad \Rightarrow \quad T(s) = \frac{y}{r_f} = \frac{G(s)C_{fb}(s)}{1 + M(s)G(s)C_{fb}(s)} \quad (4.5)$$

Including the input filter the total transfer function of the feedback loop from the reference signal  $r$  to the output  $y$  as shown in Figure 4.5 is given by:

$$T(s) = \frac{y}{r} = \frac{G(s)C_{fb}(s)}{1 + M(s)G(s)C_{fb}(s)}F(s) \quad (4.6)$$

<sup>2</sup>In many books on control theory and also a bit further in this book the output  $y$  is often assumed to include the measurement sensor as part of the plant and in that case the output disturbance and sensor disturbance are combined. In real systems with a not ideally linear sensor it is sometimes better to show the dynamic properties and limitations of the sensor as a separate transfer function.

In control design one has the freedom to choose  $F(s)$  and particularly  $C_{fb}(s)$  such that the total transfer function fulfils the desired specifications. In the pure error-feedback case, like with the positioning of the optical pick-up unit of the CD player, the reference is constant and no pre-filter is needed. In the next section, it is shown with this example, how the feedback controller  $C_{fb}(s)$  can be designed to shape the dynamics of the feedback loop.

Just as the feedforward control approach has its advantages and disadvantages, also feedback control offers some benefits and potential pitfalls:

- **Stabilisation of unstable systems:** Not all motion systems are inherently stable, some of them are marginally stable, some even unstable, like an inverted pendulum. As feedback control enables to determine the place of the closed-loop poles of the controlled system, unstable poles can be stabilised.
- **Reduction of the effect of disturbances:** Disturbances of the controlled motion system are observed in the sensor signal, and therefore the feedback controller can compensate for them.
- **Handling of uncertainties:** Feedback controlled systems can also be designed for *robustness* which means that the stability and performance requirements are guaranteed even for parameter variations of the controlled mechatronic system.

Although feedback control provides some very good features, it has of course also some pitfalls that have to be dealt with:

- **A sensor is required:** The feedback loop is closed, based on information from a sensor. Therefore feedback control only can be as good as the quality of the sensor signal allows. In precision positioning systems accurate sensors are required with high resolution and bandwidth, which are very costly. The measurement and sensing system often takes a substantial part of the total systems budget.
- **Limited reaction speed:** A feedback controller only reacts on differences between the reference signal and the measured system status, which means that the error has to occur first before the controller can correct for it.
- **Feedback of noise:** By closing the loop, the positioning noise of the motion system as well as sensor noise are also fed back, which has to be considered at the system and control design.

- **Can introduce instability:** Just as feedback control can stabilise an unstable system, it can also make a system unstable that even would be stable without control.

Feedback control is a very useful principle in motion systems as it allows to stabilise marginally stable or unstable systems. Further it introduces active damping of resonant modes, and it improves the robustness of the controlled system.

## 4.2 Feedforward control

For systems that are open-loop stable it is possible to apply feedforward control to improve the system performance when following a predefined trajectory like a reference signal or a repeating scanning motion. It was indicated in the previous section that a feedforward controller basically consists of a filter that is placed in series with the plant in order to compensate its dynamics.

### 4.2.1 Model based open-loop control

In the following an example of a *model-based* feedforward controller is introduced, a scanning unit with a piezoelectric actuator. This unit is applied as a precision positioning stage with nanometre resolution at a small range of movement. The measured frequency-response of this scanning unit is shown in Figure 4.6.

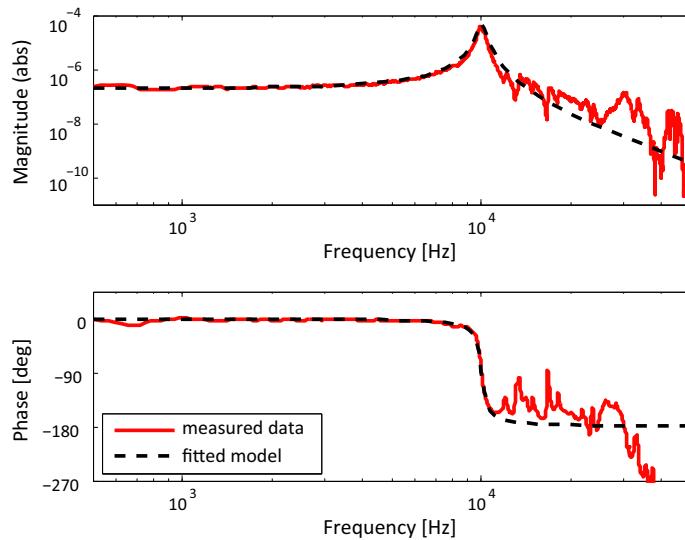
A mathematical model of a second-order mass-spring system is fitted to this measured response. The first eigenmode shows a weakly damped resonance at the eigenfrequency of 10 kHz. The following data are obtained by this system identification exercise. The natural angular frequency of the first eigenmode  $\omega_1 = 2\pi f_1 = 6.28 \cdot 10^4$  rad/s,  $\zeta_f$  as the fitted damping ratio and  $C_f$  as the fitted compliance. With these data Equation (3.32) of the previous chapter gives the transfer function of this scanning unit:

$$G(s) = \frac{C_f}{\frac{s^2}{\omega_1^2} + 2\zeta_f \frac{s}{\omega_1} + 1} = \frac{C_f \omega_1^2}{s^2 + 2\zeta_f \omega_1 s + \omega_1^2}. \quad (4.7)$$

When positioning at high speeds, the resonance at the first mode-shape of this scanning unit causes oscillations that adversely affect the tracking accuracy.

In order to solve this unfavourable behaviour a feedforward controller is defined that compensates the dynamics of this scanner by first inverting the transfer function, without changing the static gain of the positioning system. This means that the transfer function of the controller ideally would become equal to:

$$C_{ff}(s) = \frac{s^2 + 2\zeta_f \omega_1 s + \omega_1^2}{\omega_1^2} \quad (4.8)$$



**Figure 4.6:** Bode-plot of a piezoelectric-actuator based scanning unit for nanometre-resolution positioning. It shows the measured response (solid line) and the second-order model that is fitted for the low-frequency system behaviour including the resonance peak, corresponding to the natural frequency of the first mode-shape (dashed line).

This controller has a pair of zeros, corresponding with an anti-resonance at the eigenfrequency of the first eigenmode of the scanner, with equal damping. Unfortunately these zeros also imply an increasing magnitude of the transfer function controller at higher frequencies with a +2 slope in the Bode-plot. Such a behaviour is physically impossible so the controller needs to be modified in such a way that it becomes realisable. In this case it is decided to create a resulting overall transfer function of the controller and the plant that acts like a well damped mass-spring system with the same natural frequency as the plant and an additional reduction of the excitation of higher frequency eigenmodes.

In order to realise this controller first two poles have to be added, placed at the same frequency as the resonance but with a higher damping ratio. Typically a damping ratio between aperiodic and critical ( $0.7 < \zeta < 1$ ) is applied to avoid oscillations. For  $\zeta = 1$  this results in the following transfer function:

$$C_{ff}(s) = \frac{s^2 + 2\zeta_f \omega_1 s + \omega_1^2}{s^2 + 2 \cdot 1 \cdot \omega_1 s + \omega_1^2}. \quad (4.9)$$

This feedforward controller is basically a *notch filter*, because of the notch shape of the anti-resonance in the Bode-plot.

In order to create an additional attenuation of higher frequency resonances due to flexible-body mode-shapes, another first-order pole is added at the first eigenfrequency which gives the transfer function of the complete feedforward controller:

$$C_{ff}(s) = \frac{s^2 + 2\zeta_f \omega_1 s + \omega_1^2}{(s + \omega_1)(s^2 + 2\omega_1 s + \omega_1^2)}. \quad (4.10)$$

When this controller is connected in series with the scanning unit, the anti-resonance of the controller and the resonance of the piezo-scanner cancel each other out. This pole-zero cancellation is the only manipulation on poles that can be achieved with feedforward control and results in this example in a well damped third order transfer function:

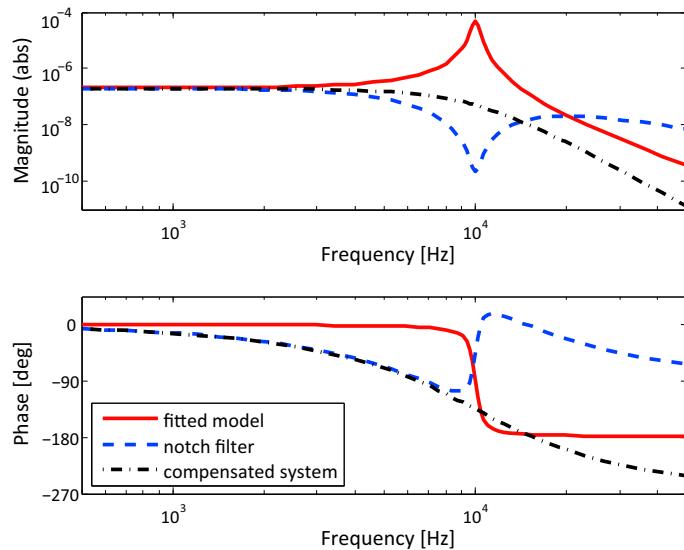
$$G_{t,ff}(s) = G(s)C_{ff}(s) \quad (4.11)$$

$$= \frac{C_f}{\cancel{(s^2 + 2\zeta_f \omega_1 s + \omega_1^2)}} \frac{\cancel{(s^2 + 2\zeta_f \omega_1 s + \omega_1^2)}}{(s + \omega_1)(s^2 + 2\omega_1 s + \omega_1^2)} \quad (4.12)$$

$$= \frac{C_f}{(s + \omega_1)(s^2 + 2\omega_1 s + \omega_1^2)} \quad (4.13)$$

The Bode-plot of the resulting mechatronic system is shown in Figure 4.7. It demonstrates the compensation of the resonance at the first mode-shape of the scanner. The controlled mechatronic system has low-pass characteristics rolling off at the scanner's first natural frequency.

The beneficial effect of such a feedforward controller on the performance of this scanner is shown in Figure 4.8. By observing the difference between the performance of the open-loop controlled and uncontrolled scanning unit in a triangular scanning motion at 1000 lines per second, it is clear that the scanner oscillations are suppressed and also the tracking of the triangular scanning signal is improved significantly.



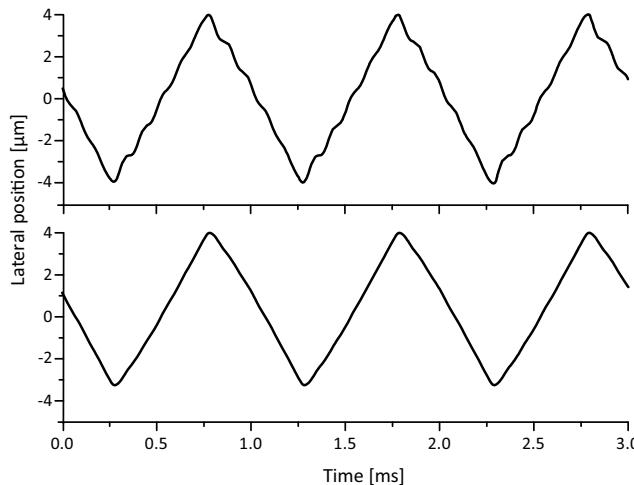
**Figure 4.7:** Bode-plot of a feedforward-controlled scanning unit for nanometre-resolution positioning. It shows both the fitted dynamic model of the scanning unit without control (solid line), the notch filter by the 3<sup>rd</sup>-order feedforward controller (dashed line) and the resulting compensated dynamic performance of the combined scanning unit and controller (dashed-dotted line).

### 4.2.2 Input-shaping

Another open-loop method, that is often used in motion control, is called *input-shaping*. With this method the reference signal is modified in a different way than by the linear filtering and compensation as shown in the previous section. As an example of this method, the comparable piezoelectric-actuator driven scanning unit for nanometre-resolution positioning is used. In this case however, the system dynamics of the positioning system are dominated by a resonance frequency occurring at 22 kHz.

When applying a step signal (in simulation) to the scanning unit, it would start to oscillate at its natural frequency where the oscillations would fade away after the step according to the damping of this resonance. In a first approximation, the scanner can be assumed to behave like a linear system which means that a reduction of the input step stimulus by a factor of two would result in a reduction of the amplitude of the response by the same factor two.

In case two of these steps are applied with only half the height of the full

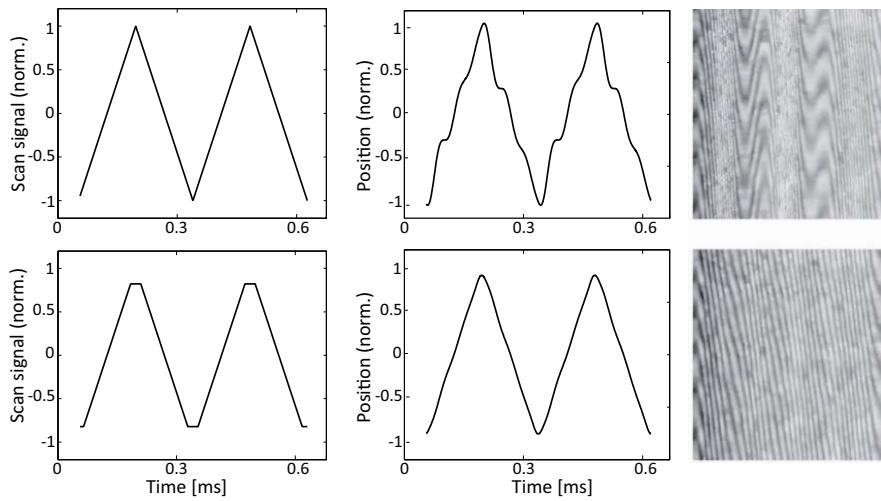


**Figure 4.8:** Laser-Vibrometer measured scanning motion of the piezoelectric-actuator driven positioning system in response to a triangular scanning signal at 1000 lines per second. Without control (upper graph) oscillations of the scanning unit are clearly observed at the sharp transitions in the scanning signal. With pole-zero cancellation based open-loop control (lower graph) these oscillations are suppressed and the tracking accuracy is improved.

step, the same steady-state response would be obtained as with the full step stimulus after all oscillations are damped out. If one of these half-height steps is delayed by half the period of the scanner's resonance frequency, the oscillations that are caused by each individual step are  $180^\circ$  out of phase, and cancel each other out.

This splitting of the reference signal into two (equal) signals and delaying one of them by half the period of the system's resonance is a typical example of input-shaping. This method clearly is very different from pole-zero cancellation as it is time domain instead of frequency domain based filtering. In the frequency domain these sampled adaptations to the input create a frequency spectrum with a multiple of notch filters at the harmonics of the frequency that these adaptations are applied. This effect is related to real sampling that will be explained in Chapter 8 on measuring.

With the previous example of the feedforward controller with pole-zero cancellation it was shown that a triangular scanning signal to the scanning unit "triggers" oscillations after each change of slope of the triangular signal. These are caused by the discontinuity in the derivative and the corresponding higher harmonics of the triangle waveform as presented by Fourier



**Figure 4.9:** Input-shaping control of the triangular scanning signal in a scanning-probe microscope at 3900 lines per second. Due to the shaping of the input signal by cutting the top of the triangular signal, oscillations of the moving part are significantly reduced. The left column shows the input signals without (top) and with (bottom) input-shaping while the middle column shows the resulting motion profiles of the scanner. The pictures on the right show a set of AFM data of a parallel grating that has a pitch of 233 nm. The distortion and reduction of distortion of the grating structure is clearly visible.

analysis in Chapter 2.

Applying input-shaping to this triangular scanning signal results in the introduction of a plateau instead of the sharp peak, where the width of the plateau corresponds to half the period of the scanner's resonance as can be seen in Figure 4.9. When at a positive slope the scanning signal reaches the plateau and stops raising, the scanner starts to oscillate at its natural frequency, so with a sinusoidal motion. After half the period of this oscillation, the scanner is arrived at the same position as where the oscillation started, but it is moving with almost the same speed in the opposite direction, due to the low damping. When at that moment the scanning signal is changed into a negative slope, a smooth transition is realised, without further exciting the scanner's resonance, and the actual motion follows the desired triangular scanning motion.

### 4.2.3 Adaptive feedforward control

It is important to emphasise that both examples of feedforward control, the model-based pole-zero cancellation and the input-shaping, only work reliably as long as the dynamic properties of the total plant are known and remain constant. These dynamics include the transfer functions of passive elements like the mechanics as well as active elements like the amplifiers and actuators. In reality often external influences have an impact on these dynamic properties, leading to an increasing deviation between the parameters in the model and the reality. This deviation can be partly solved by *adaptive feedforward control*, adapting the feedforward signal by measuring its real behaviour. This method requires a sensor to obtain information about the behaviour and for that reason it is often applied in combination with feedback.

For repetitive processes such as a scanning motion the residual tracking error can be learned from previous scans and can be used to optimise the open-loop scanning signal that is applied to the motion system. This version of adaptive feedforward control is called *iterative learning control*. By measuring the output of a motion system over the known trajectory and comparing the measured behaviour with the intended behaviour, information is gathered about the amount of error in the model and the direction of the necessary correction.

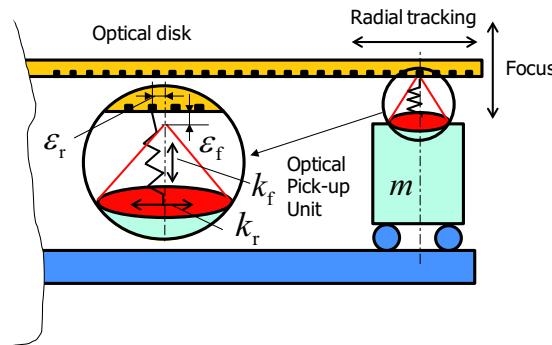
### 4.3 PID feedback control

In the previous section it was shown how feedforward control can work for an open-loop stable system with known properties and circumstances. Striving for a maximum of predictability of the dynamic behaviour of a mechatronic system is an important part of the design because of the benefits of feedforward control. In spite of these efforts, even with the most sophisticated models and by suppressing external disturbances to the bare minimum, almost always some remaining errors need to be corrected by feedback control. Next to these systems with remaining random errors, motion systems with inherent instability, like a magnetic bearing, with poles in the right-half-plane require feedback control for stabilisation.

Feedback control is more complex and critical to design than feedforward control. For that reason it receives generally more attention in the scientific world and in the last decades several approaches have been introduced to design an “optimal” feedback control system according to several optimisation criteria. In most cases these innovations were driven by the new possibilities of fast computer systems that enabled “real-time” modelling and control of higher-order systems. Still the classical *PID-control* principle, where PID stands for a proportional-integral-differential feedback loop is predominantly used in industry because of its relatively straightforward mode of operation combined with generally a sufficient performance. Knowledge about the different properties of PID-control is crucial for the design of high performance mechatronic systems. Without mastering this knowledge all further refinements in control theory are of no practical value. For this reason this section is fully devoted to PID-control with a full emphasis on its representation in the frequency domain.

In spite of this dominance of PID-control in mechatronic systems it is expected that modern control methods will gradually be used more frequently. For that reason some of the promising new elements of modern control engineering will be introduced after this section, based on the *state-space* representation in the time domain.

Before explaining the more generic methodology, PD-feedback control, shortly named *PD-control* of an optical pick-up unit for a compact disk player is presented to create an initial idea about the principle. This unit was introduced in Chapter 3 on dynamics. In control terms this is a *servo-system*. The word “servo” stems from the Latin word “servus”, which means “slave” and refers to a control system that is designed to follow a moving target.



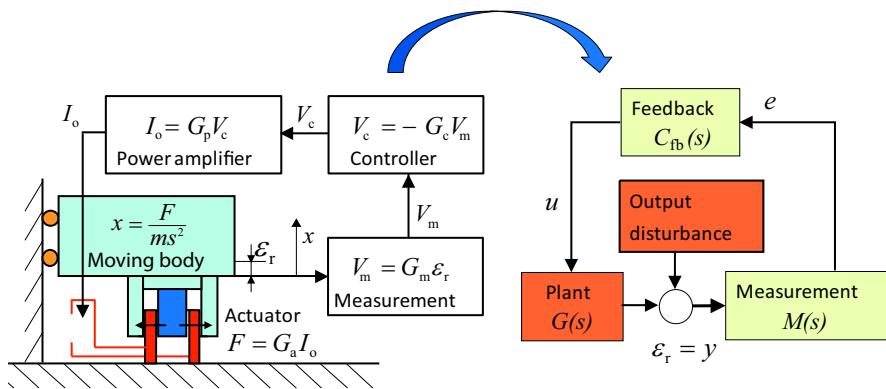
**Figure 4.10:** The optical disk pick-up unit of Chapter 3 connected to the track by means of a “virtual” spring, created by the position control system.

### 4.3.1 PD-control of a Compact-Disc player

In the introduction of the optical pick-up unit for a compact disk player the relation between the control task and the maximum position error of the lens in respect to the track of the CD was explained. The control system creates a virtual connection between the track and the optical system as indicated in Figure 4.10 by a mechanical spring. It was shown that the required radial stiffness  $k_r$  amounts to a value of  $2.5 \cdot 10^5 \text{ N/m}$  for a moving mass  $m = 10^{-2} \text{ kg}$  and a corresponding natural frequency  $f_0$  of 800 Hz. This stiffness appeared to be equal to the total loop gain  $G_t$  of the feedback loop around the moving body that consisted of the combination of the gain of the measurement system, the controller, the power amplifier and the actuator, as shown schematically at the left side of Figure 4.11. It was further explained that a spring and a mass need a damper to control the resonance at its natural frequency. For this reason in this section also a virtual damper will be added.

At the right side of Figure 4.11 the equivalent feedback loop for the position control mechanism is shown. It is a simplified version of Figure 4.5 without a reference signal and pre-filter. The plant consists of the amplifier, the actuator, the mechanical moving body and the measurement sensor.

The loop will be traced “step by step” from within the plant. The input of the mechanics is the force that is generated by the actuator of the positioning system that holds the lens. The output of the (already perturbed) motion system is the radial tracking position error  $\varepsilon_r$  between the lens position and the centre of the track. In the equivalent feedback loop this radial error is denoted by the control error  $e$  after the measurement.



**Figure 4.11:** Mass-spring system with a virtual spring that is created by closed-loop control. The plant consists of the power amplifier, actuator, moving body and measurement system. In combination with the controller the total system can be described by the simple feedback loop as shown at the right side of the drawing. The disturbance, consisting of the movement of the track due to eccentricity and external vibrations, directly interferes with the output of the plant.

The positioning system with the lens is modelled as a single body with mass  $m$  and its corresponding compliance transfer function  $C_m(s)$  is equal to:

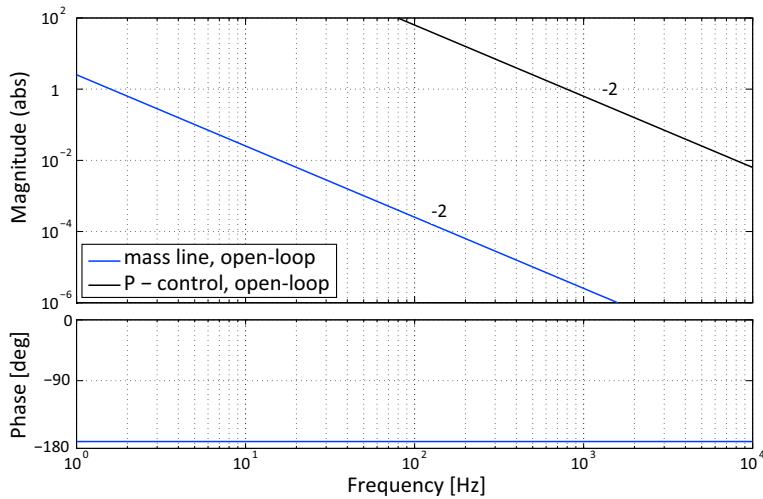
$$C_m(s) = \frac{x}{F} = \frac{1}{ms^2} = -\frac{1}{0.01\omega^2} \quad (4.14)$$

The transfer function shows a  $-2$  slope in the Bode-plot and a phase lag of  $180^\circ$  over the entire spectrum. The magnitude of the compliance shows a gain of one at a *unity-gain cross-over frequency* of  $\approx 1,6$  Hz, as shown in the lower (blue) mass-line of Figure 4.12.

#### 4.3.1.1 Proportional feedback

As a first step, the previously shown stiffness that is created by feedback will be described in the common terms of control engineering. For this example the gains of the amplifier, actuator and measurement sensor are all assumed to be frequency independent constant factors with a value of one for reason of simplicity. This means that the transfer function  $G(s)$  of the total plant is equal to the compliance of the moving body  $C_m(s)$ .

It was shown in Chapter 3 that the total loop gain  $G_t$ , excluding the transfer function of the body, has to be equal to the required radial stiffness  $k_r = 2.5 \cdot 10^5$ . This means that the controller needs a *proportional* gain of  $k_p = G_t =$



**Figure 4.12:** Open-loop Bode-plot of the CD-player lens with a mass of  $10^{-2}$  kg. P-control with a gain  $k_p = 2.5 \cdot 10^5$  proportionally shifts the total frequency response of the combined system upwards, such that the unity-gain cross-over frequency becomes equal to 800 Hz.

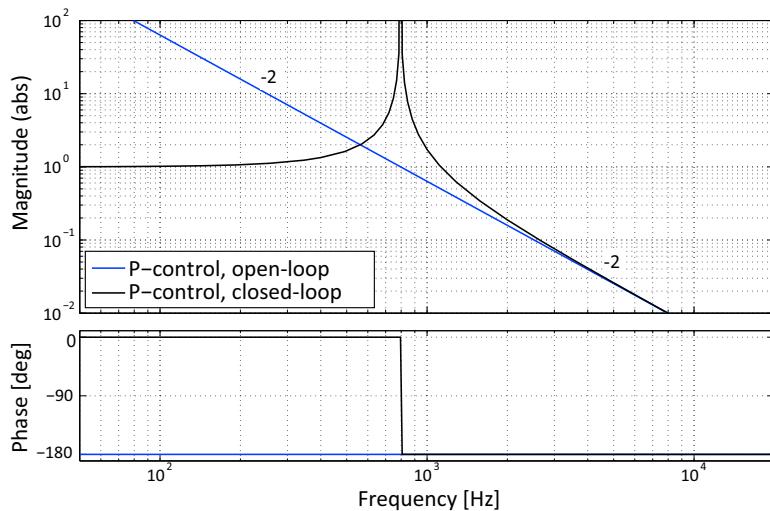
$2.5 \cdot 10^5$ . The related control term of this proportional gain is *proportional control* or *P-control*. In the loop this means that the feedback control transfer function  $C_{fb}(s)$  becomes  $C_p = k_p$ , without the Laplace variable as there is no frequency dependency of the gain value.

Before the loop is closed, the *open-loop* Bode-plot of the P-control feedback system is shown in the upper (black) mass-line of Figure 4.12. The open-loop transfer function is the total transfer function of the system from the input of the controller following all functional elements until the measured sensor signal  $e$ , without connecting this signal with the feedback input of the controller. This open-loop transfer function  $L_p(s)$  in the frequency domain is equal to:

$$L_p(s) = G(s)C_p = C_m(s)C_p = \frac{k_p}{ms^2} \quad \Rightarrow \quad L_p(\omega) = -\frac{2.5 \cdot 10^5}{0.01\omega^2}. \quad (4.15)$$

P-control shifts the mass-line of the lens-system upwards in the Bode-plot because it is only a simple gain factor. This shift in magnitude also causes a shift of the unity-gain cross-over frequency ( $\omega_c = 2\pi f_c$ ) to a higher level, corresponding with the previously determined required value of the natural<sup>3</sup> frequency  $f_0 = 800$  Hz, which was directly related to the required maximum

<sup>3</sup>The natural frequency  $\omega_0 = 2\pi f_0$  relates primarily to the first resonance frequency of a passive dynamic mass-spring system. Later in this chapter an example is shown with a passive



**Figure 4.13:** Open-loop and closed-loop Bode-plot of the feedback controlled CD-player lens with P-control at a gain of  $k_p = 2.5 \cdot 10^5$ . The resulting undamped resonance at the natural frequency of 800 Hz is clearly visible.

position error.

When the feedback loop with only P-control is closed, the transfer function becomes in the frequency domain:

$$T_p(s) = \frac{L_p(s)}{1+L_p(s)} = \frac{1}{\frac{m}{k_p}s^2 + 1} \quad \Rightarrow \quad T_p(\omega) = \frac{1}{-\left(\frac{\omega}{\omega_c}\right)^2 + 1}. \quad (4.16)$$

with the corresponding Bode-plot of the closed-loop system as shown in Figure 4.13.

When the open-loop and the closed-loop response are compared, the first difference is found at the frequency range below the unity-gain cross-over frequency of the open-loop. The closed-loop response of the system shows a constant gain of one, corresponding with a spring-line, and its phase shift has become zero degrees. At the cross-over frequency a sharp resonance peak is observed, and above this frequency the original  $-2$  slope of the open-loop response is followed.

When comparing the Bode-plot of the closed-loop system to the passive dynamic systems discussed in the previous chapter, it is fully reconfirmed that proportional feedback control creates a spring action.

---

dynamic system with resonance frequency  $\omega_0$ , where the unity-gain cross-over frequency  $\omega_c$  is different from  $\omega_0$ . For that reason the different terms will be used throughout this chapter.

The resonance peak is of course not acceptable for the closed-loop operated system and in control engineering terms this effect is represented by the location of the poles of the system in the Laplace plane. The closed-loop poles are a conjugate complex pair on the imaginary axis which means that the system is marginally stable as the phase-lag of the open-loop transfer function at the cross-over frequency is  $180^\circ$ . In order to achieve an acceptable control performance, some damping and phase-lead around the unity-gain cross-over frequency needs to be added.

#### 4.3.1.2 Proportional-differential feedback

In order to add damping to the positioning system of the lens, a force component must be created, that is proportional to the velocity. This force can be created by a signal that corresponds with the derivative of the position error over time. With this derivative signal the feedback controller transfer function  $C_{fb}(s)$  becomes equal to the *Proportional Differential control* or PD-control transfer function  $C_{pd}(s)$ :

$$C_{fb}(s) = C_{pd}(s) = k_p + k_d s \quad (4.17)$$

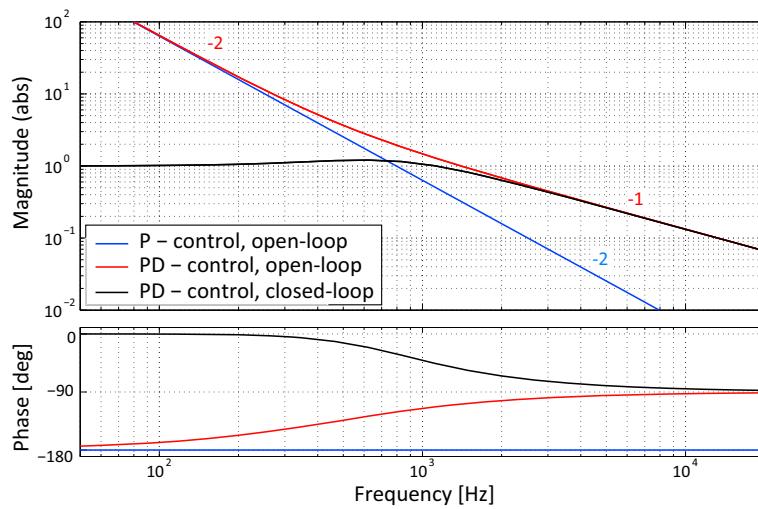
This PD-control transfer function is also often written in a different form:

$$C_{pd}(s) = k_p \left(1 + \frac{k_d}{k_p} s\right) = k_p (1 + \tau_d s) \quad (4.18)$$

In this second notation  $\tau_d$  represents the time constant of a first-order differentiator that increases the magnitude of the transfer function with a  $+1$  slope above the radial frequency  $\omega_d = 1/\tau_d$ . For radial frequencies below  $\omega_d$  the derivative term ( $s$ ) is smaller than one and the feedback controller behaves like a P-control system. For frequencies that are higher than this ratio the derivative term will dominate and the controller shows mainly differentiating behaviour.

In order to provide the desired damping effect, this change from proportional characteristic to differentiating action has to take place at a sufficiently lower frequency than the unity-gain cross-over frequency of the open-loop transfer function  $L(s)$ . In a mass-line based positioning system, as a “rule of thumb” a good performance typically is achieved when the differentiating action starts at one third of the unity-gain cross-over frequency. The combined open-loop transfer function of the PD-control with the mass response of the CD player lens equals:

$$L_{pd}(s) = G(s)C_{pd}(s) = \frac{k_p + k_d s}{m s^2}. \quad (4.19)$$



**Figure 4.14:** Bode-plot of the response of the feedback controlled CD player lens with PD control without limitation of the differentiating action at high frequencies. The unity-gain cross-over frequency is increased and the closed-loop response has a  $-1$  slope at high frequencies, which implies a reduced attenuation of noise and resonances due to high-frequency mode-shapes.

In Figure 4.14 the response of this transfer function is shown in red with a  $-2$  slope at the frequency range where  $k_p$  is dominating. In the range where the differentiation dominates, a  $-1$  slope is visible in the amplitude-plot with a corresponding less negative phase in the frequency range of the unity-gain cross-over frequency. The resulting phase lag is approximately  $120^\circ$  as compared to the original  $180^\circ$ . For this reason a differentiating action in the control system is also often called a *lead-network* as it introduces a phase lead in the open-loop response, reducing the original phase lag of the mass response at the unity-gain cross-over frequency.

Closing the feedback loop for this system results in the following transfer function of the closed-loop system:

$$T_{pd}(s) = \frac{L_{pd}(s)}{1 + L_{pd}(s)} = \frac{k_p + k_d s}{m s^2 + k_p + k_d s} \quad (4.20)$$

This equation can be written in the frequency domain in the form as presented in the previous chapter for passive dynamic systems with the ratio

$\omega/\omega_c$  as variable:

$$T_{pd}(s) = \frac{\frac{k_d}{k_p}s + 1}{\frac{ms^2}{k_p} + \frac{k_d}{k_p}s + 1} = \frac{\frac{s}{\omega_d} + 1}{\frac{s^2}{\omega_c^2} + \frac{s}{\omega_d} + 1} \quad (4.21)$$

$$T_{pd}(\omega) = \frac{2j\zeta\frac{\omega}{\omega_c} + 1}{-\left(\frac{\omega}{\omega_c}\right)^2 + 2j\zeta\frac{\omega}{\omega_c} + 1} \quad (4.22)$$

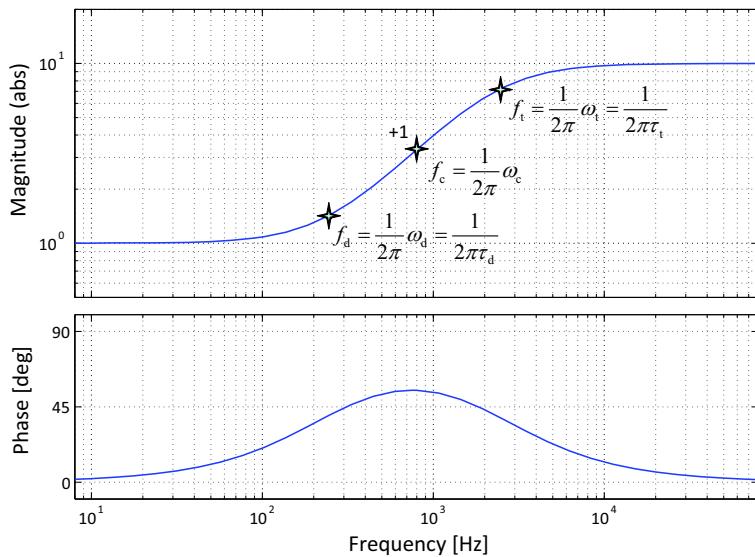
with  $\omega_c = 2\pi f_c = \sqrt{k_p/m}$  and  $\omega_d = k_p/k_d = 1/\tau_d = \omega_0/2\zeta$ .

This transfer function corresponds with the characteristic transfer function of the mechanical damped transmissibility of a movement of the support to a body, connected by a spring and a damper as expressed in Equation 3.64 in the previous chapter. The change of the  $-2$  slope to a  $-1$  slope at higher frequencies is caused by the zero in the numerator of the transfer function. This result is not surprising as the optical lens from the CD player is also following the movement of the track, but in this case it is connected by a virtual spring and a virtual damper. Likewise with the transmissibility example, this configuration has a drawback due to the dominant coupling of the damper at higher frequencies but in active feedback this effect can be reduced by limiting the virtual damping at these frequencies.

#### 4.3.1.3 Limiting the differentiating action

In a real positioning system many resonating mode-shapes occur at higher frequencies and it is better to attenuate these to prevent instability. It is also physically impossible and would not be wise to create a controller with an infinite gain at infinite frequencies. For the active PD-controlled system this means that the differentiating action must be limited at higher frequencies, sufficiently above the unity-gain cross-over frequency in order to retain the beneficial phase lead.

As a “rule of thumb”, typically a frequency of three times the cross-over frequency is chosen as the frequency above which the differentiating action is terminated. This *tamed PD-control* results in a characteristic behaviour around the cross-over frequency that is sometimes also called *lead-lag compensation* because the differentiation introduces a zero which gives a phase lead and the termination is done by adding a pole which gives a phase lag, compensating the lead of the differentiator zero. This compensator does not really induce a lag as after the termination of the differentiation, the phase



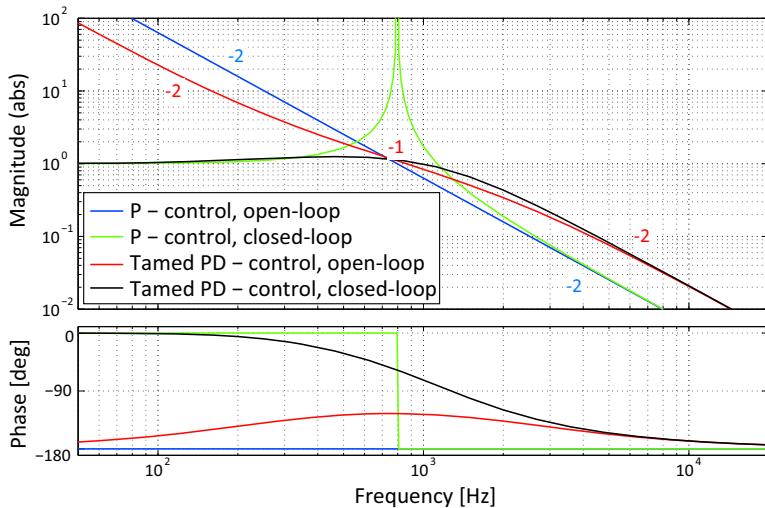
**Figure 4.15:** Bode-plot of limited or tamed D-control. The differentiating action starts at  $f_d$  and is terminated at  $f_t$ . With the “rule of thumb”  $f_d = 0.33f_0 = 0.1f_t$  a phase lead of approximately  $55^\circ$  is obtained in the controller.

of the combined controller becomes zero and the gain becomes proportional again. This is shown in Figure 4.15 where  $f_d = \omega_d/2\pi$  and  $f_t = \omega_t/2\pi$ . It will be shown later that an additional low-pass filter at  $f_t$  is in practice even more preferred in which case the slope becomes  $-1$  and a real roll-off with phase lag is obtained at higher frequencies.

Another issue of the PD-control action can also be observed in Figure 4.14 as the cross-over frequency is increased significantly due to the magnitude increase by the differentiating term. With the typical “rule of thumb” value  $\omega_d = 0.33\omega_c$  it is necessary to reduce  $k_p$  with the same factor of 0.33 to achieve an unchanged cross-over point. With that correction, the total open-loop transfer function of the system becomes in the frequency domain:

$$L_{\text{pdt}}(s) = G(s)C_{\text{pdt}}(s) = \frac{\omega_d}{\omega_c} \frac{k_p}{ms^2} \frac{1 + \frac{1}{\omega_d}s}{1 + \frac{1}{\omega_t}s} = \frac{\omega_d}{\omega_c} \frac{k_p}{ms^2} \frac{1 + \tau_d s}{1 + \tau_t s}. \quad (4.23)$$

with  $\omega_t = 1/\tau_t$  being the radial frequency where the differentiator is “tamed” at approximately three times  $\omega_c$  according to the “rule of thumb” which is also approximately ten times  $\omega_d$ . The Bode-plot of the resulting response with  $\omega_d = 0.33\omega_c = 0.1\omega_t$  or  $f_d = 0.33f_0 = 0.1f_t$  is shown in Figure 4.16 both

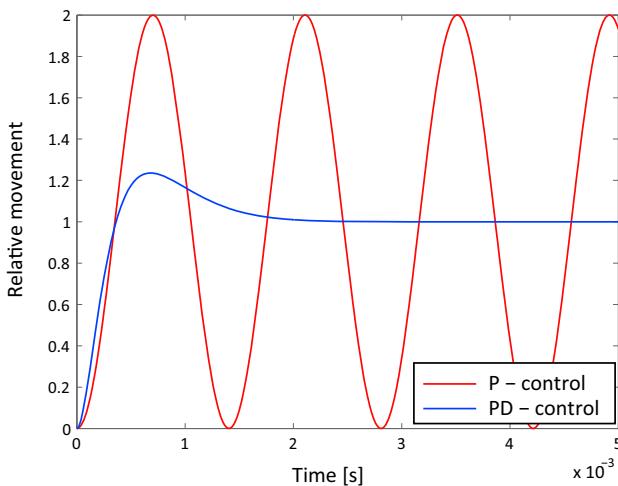


**Figure 4.16:** Open-loop and closed-loop Bode-plot of the feedback controlled CD-player lens for both P- and tamed PD-control with  $\omega_d = 0.33\omega_c = 0.1\omega_t$ . It shows the reduction of the loop gain at lower frequencies and the reduction of the attenuation at higher frequencies due to the necessary phase compensation at the unity-gain cross-over frequency.

in open- and closed-loop, for comparison together with the response with P-control only.

The unity-gain cross-over frequency of the open-loop transfer function has become the *roll-off frequency* of the closed-loop system. The closed-loop frequency response shows a nicely damped behaviour at the roll-off frequency with a  $-2$  slope, but at a sacrifice. First the total loop gain at low frequencies is reduced by a factor 3, corresponding to a proportional reduction of the virtual stiffness. As a direct consequence also the capability to suppress errors at those frequencies is reduced. Secondly also at high frequencies the attenuation of resonances and noise is reduced with a factor 3. It is important to notice that this is an area of optimisation as  $\omega_d$  and  $\omega_t$  can be placed closer together to achieve better error reduction at low and high frequencies with a reduced damping around the unity gain cross-over frequency  $\omega_c$ .

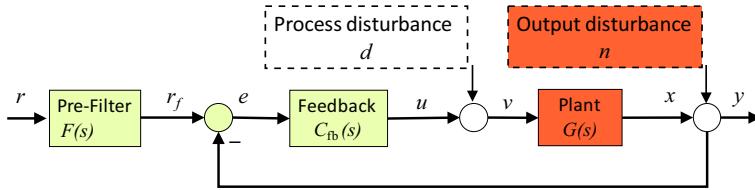
As another method to solve the reduction of the loop gain at lower frequencies, an integrating control (I-control) can be introduced eventually forming the well-known *PID-control* principle. First some more control theory will be presented before further introducing this useful principle, because I-control has no simple comparable element in the mechanical world.



**Figure 4.17:** Step response of the closed-loop controlled CD-player lens under PD-control and P-control, clearly showing the beneficial damping effect on the resonance at the unity-gain cross-over frequency  $\omega_c$ .

To finalise this example the simulated response of the closed-loop system to a step in the reference signal is illustrated in Figure 4.17, clearly showing that the PD-controlled system nicely follows the step in the reference signal with only a small overshoot, whereas the pure P-controlled system continuously oscillates around the final value due to the absence of damping.

It is worthwhile repeating here, that although *simulated* step responses may be a useful tool to evaluate the control performance of a mechatronic system, special care has to be taken, when eventually testing the controller on the real hardware. In that situation step or even impulse like reference signals may result in too aggressive peaks in the control signal and consequently in the actuation force. Smoother signals like sinusoidal or low-pass filtered steps and ramp-like reference signals may be more appropriate for testing of the control performance on the real system. Of course care also has to be taken when choosing the amplitude of the test signals in order not to get into trouble with non-linearity or even damage to the hardware.



**Figure 4.18:** Simplified representation of a feedback loop in order to determine the influence of the reference signal and the most important disturbance sources on the outputs of the plant  $x$ , the feedback controller  $u$  and the complete system  $y$ . The reference and feedback control input are in the same position domain as the output of the complete system. The process disturbance includes all disturbances from within the plant while the output disturbance includes also the measurement errors.

### 4.3.2 Sensitivity functions of feedback control

As was demonstrated in the example of the tracking controller for the CD-player lens, the dynamics of the controlled system can be directly modified by closing the feedback loop. Feedback control allows to directly place the system poles at values that are more useful for the operation of the motion system than their natural locations. This enables a faster response of the system with adequate damping. This property is in clear contrast to feedforward control, where only pole-zero cancellation can be applied by compensation with the inverse transfer function.

One of the other main advantages of feedback control was the possibility to reduce the effect of disturbances. To illustrate this a simplified version of the generic feedback loop is shown in Figure 4.18. The measurement system is assumed to have a unity-gain transfer function. This means that for this thinking model the reference and the input of the controller work in the same position domain as the output of the system. The disturbance in the measurement system is included in the output disturbance  $n$ . The process disturbance can be anything occurring between the output of the feedback controller and the output of the plant. For reason of simplicity it is inserted between the plant and the controller. Disturbances from within the plant can be calculated to this location by using the mathematical model of the plant. With this simplified model, the transfer functions of the different inputs of the system to three relevant output variables in the loop are written down in a set of equations. The input variables include the reference and the two sources of disturbances. The first output variable is  $x$  for the output of the plant, being a subset of the system-states that will be introduced in

Section 4.4. The second and third output variables are  $y$  for the output of the total system and  $u$  for the output of the controller. In these transfer functions for simplicity the Laplace variable ( $s$ ) is omitted and  $C = C_{fb}$ :

$$x = \frac{G}{1+GC}d - \frac{GC}{1+GC}n + \frac{GCF}{1+GC}r \quad (4.24)$$

$$y = \frac{G}{1+GC}d + \frac{1}{1+GC}n + \frac{GCF}{1+GC}r \quad (4.25)$$

$$u = -\frac{GC}{1+GC}d - \frac{C}{1+GC}n + \frac{CF}{1+GC}r \quad (4.26)$$

It should be noted that these transfer functions are written in a purely single dimensional notation with scalar values. When more degrees of freedom are controlled the different upper-case letters become matrices and the lower-case letters become vectors. With multiplication of matrices the successive order is important and the shown order is valid for the defined feedback loop from Figure 4.18. When for instance disturbance  $d$  is neglected, the vector notation would be  $\mathbf{x} = \mathbf{Gv}$  and  $\mathbf{v} = \mathbf{Ce}$  which gives  $\mathbf{x} = \mathbf{GCe}$ .

Although these transfer functions consist of a combination of nine transfer functions, one for each input-output combination, some of them are the same. When these doubles are omitted, six different transfer functions remain as shown in Table 4.1. These functions are also called the *gang of six*, a term introduced by the Swedish control scientist Karl Johan Åström from Lund University, because this set of transfer functions gives an interesting insight in how the feedback controlled system reacts to the different system inputs.

The transfer functions of the first column shows the influence of the reference signal on the output of the plant and on the output of the feedback controller. The first transfer function of the second column gives the influence of the output disturbance on the output of the plant. while both transfer functions in the second column give the influence of the two disturbances on the output of the controller.

The first transfer function in the third column shows how the output of the plant reacts to the process disturbance. The second transfer function of the third column gives the influence of the output disturbance on the output of

**Table 4.1:** Gang of six.

$\frac{x}{r} = \frac{y}{r} = \frac{GCF}{1+GC}$	$-\frac{x}{n} = -\frac{u}{d} = \frac{GC}{1+GC}$	$\frac{x}{d} = \frac{y}{d} = \frac{G}{1+GC}$
$\frac{u}{r} = \frac{CF}{1+GC}$	$\frac{u}{n} = \frac{C}{1+GC}$	$\frac{y}{n} = \frac{1}{1+GC}$

**Table 4.2:** Gang of four.

$$\begin{array}{c|c} \frac{x}{r} = \frac{y}{r} = -\frac{x}{n} = -\frac{u}{d} = \frac{GC}{1+GC} & \frac{x}{d} = \frac{y}{d} = \frac{G}{1+GC} \\ \hline \frac{u}{r} = \frac{u}{n} = \frac{C}{1+GC} & \frac{y}{n} = \frac{1}{1+GC} \end{array}$$

the total system.

In case no input filter is applied  $F$  is equal to one and the first column becomes equal to the second column, reducing the gang of six to the *gang of four* as shown in Table 4.2. This short set of equations also corresponds with the situation without a reference like the pure error feedback system of the CD-player.

From these four, the two most important transfer functions are the *complementary sensitivity function*  $T(s)$ , corresponding with the first transfer function of the first column and the *sensitivity function*  $S(s)$ , corresponding with the second transfer function of the second column.

Written in full Laplace notation the complementary sensitivity function equals:

$$T(s) = \frac{y}{r} = \frac{G(s)C_{fb}(s)}{1 + G(s)C_{fb}(s)}, \quad (4.27)$$

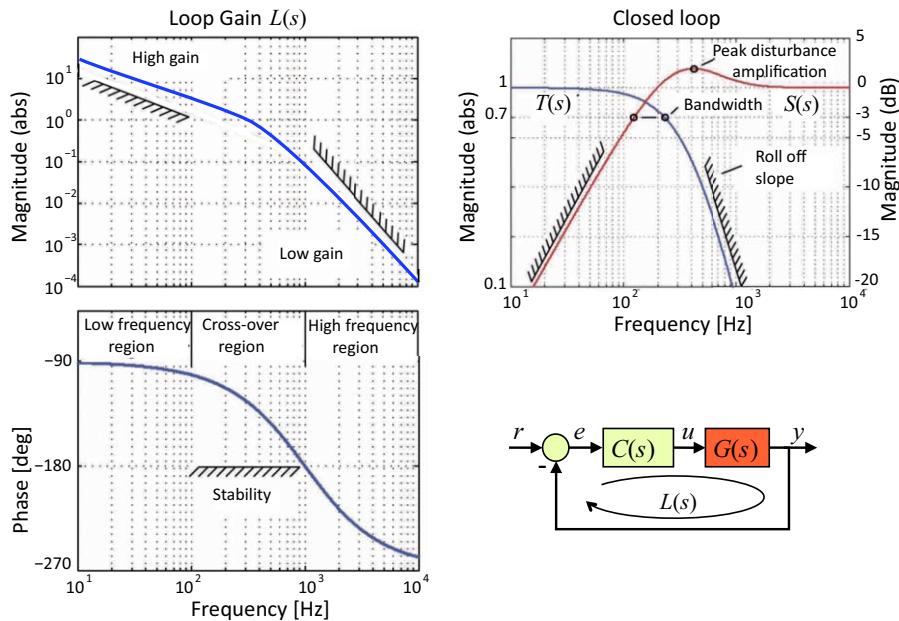
The complementary sensitivity function represents the system response to the reference in case  $F = 1$ . It gives an indication how well the system behaves as a servo-system, reliably following a target or a reference.

Written in full Laplace notation the sensitivity function equals:

$$S(s) = \frac{y}{n} = \frac{1}{1 + G(s)C_{fb}(s)} \quad (4.28)$$

The sensitivity function represents the ability of the feedback controlled system to reject disturbances.

The name "complementary" in  $T(s)$  is based on the fact that  $T(s)$  and  $S(s)$  add to unity, ( $T(s) + S(s) = 1$ ).



**Figure 4.19:** Stability condition and robustness of a feedback controlled system.

The desired shape of these curves guide the control design by optimising the levels and slopes of the amplitude Bode-plot at low and high frequencies for suppression of the disturbances and of the phase Bode-plot in the cross-over frequency region. This is called *loop shaping design*.

### 4.3.3 Stability and robustness in feedback control

In the CD player example it was demonstrated that the open-loop transfer function of the PD-control feedback system needs to have certain properties. After closing the loop, the controlled system should be stable and show good performance regarding the accuracy in following the track. Based on these requirements, certain conditions for stability and robustness to changing conditions can be derived. These conditions for stability and robustness also give insight in how the open-loop system should look like such that the closed-loop system performs well. These insights can be used to select the controller parameters for properly adjusting the transfer function of the open-loop system. The optimal tuning of a feedback loop is called *loop-shaping design*.

The most important and characteristic frequency area for the analysis of a controlled mechatronic system is around the unity-gain cross-over frequency,

as shown in Figure 4.19. In control engineering the term *bandwidth* is often used in relation to this unity-gain cross-over frequency as above this frequency the loop gain becomes smaller than one and consequently the feedback controller becomes no longer effective. Usually the term bandwidth is defined as the frequency band where the power of the output signal of a system becomes less than half the desired power level. In terms of signal amplitude the corresponding value is equal to  $1/\sqrt{2} \approx 0.7$ . In decibels this value is equal to  $-3$  dB and this value is a well-known definition for the bandwidth of filters and other frequency dependent functional devices like loudspeakers.

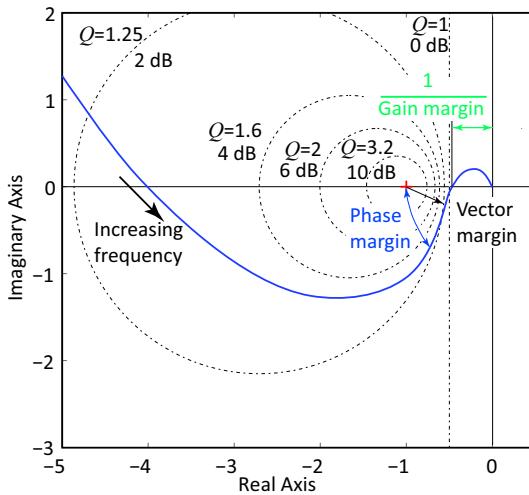
In the context of precision positioning systems it is better to define the bandwidth as the frequency range, where the amplitude of the **open-loop** transfer function exceeds a value of one, because it is the open-loop gain at certain frequencies that is important in mechatronic systems.

The condition for closed-loop stability is that the total phase-lag of the open-loop system, consisting of the feedback controller in series with the mechatronic system, must be less than  $180^\circ$  in the frequency region of the cross-over frequency. A system that has exactly  $180^\circ$  phase-lag at the cross-over point is called marginally stable, like for example the CD player with only P-control. In this situation the smallest additional time-delay or phase-lag would make the closed-loop system unstable.

To determine the stability one can also analyse the poles of the closed-loop system, where the stability condition is the same as for any linear system. For asymptotic stability all system poles have to have a strictly negative real part, which means that they have to be located in the left-half of the Laplace plane.

The Nyquist plot, like shown in Figure 4.20, is the best frequency response representation to analyse the robustness of a feedback system by the distance and direction of the graph relative to the location of the  $-1$  point on the real axis. In this graph three values are shown that relate to the robustness of the closed-loop feedback system, the *gain margin*, *phase margin* and *vector margin*.

The **gain margin** determines by which factor the open-loop gain additionally can increase before the closed-loop system goes unstable. It is defined by the distance between the loop-gain  $L(s)$  and unity-gain at the frequency where the phase-lag of  $L(s)$  becomes more negative than  $-180^\circ$ . The gain margin can have values between zero and infinite. With first and second order transfer functions where the phase does never become more negative than  $-180^\circ$  the gain can be increased



**Figure 4.20:** The phase, gain and vector margin can be easily derived in a Nyquist plot. Stability is guaranteed when the  $-1$  point on the real axis is kept at the left hand side of the response-line upon passing with increasing frequency. The dashed circles determine the magnitude peak ( $Q$ ) after closing the loop at the frequencies where the response-line crosses the circles. The blue response-line represents an example PID-controlled mass-positioning system where the I-control action together with the mass creates a phase of more than  $-180^\circ$  at low frequencies.

theoretically to infinite, corresponding with an infinite gain margin.

**The phase margin** determines how much additional phase lag at the unity-gain cross-over frequency is acceptable before the closed-loop system becomes unstable. It is defined by the difference between the actual phase-lag of  $L(s)$  and  $-180^\circ$  at the unity-gain cross-over frequency.

**The Vector margin** is defined by the closest distance in a Nyquist plot between the graph and the  $-1$ -point on the real axis.

The following rules for stability can be derived from the gain- and phase-margin in relation to the Nyquist plot:

- The system will become unstable upon closing the loop if the phase of the open-loop transfer function passes the negative real axis ( $-180^\circ$ ) at an amplitude larger than one and does not return below the negative real axis before the gain gets smaller than one.

- A stable system after closing the loop is recognised in the Nyquist plot when the  $-1$  point on the real axis is kept at the left hand side upon passing with increasing frequencies.

It is also interesting to observe in this example plot that the phase at low frequencies is more than  $-180^\circ$  shifted. This situation occurs when the I-control action, that will be presented in the following subsection, is applied with a purely mass based positioning system like the CD player pick-up unit, so without a spring to the stationary reference. Even though such a system is stable in the closed-loop situation, it is not readily known what the amount of damping will be when the feedback loop is closed. To indicate the magnitude of the closed-loop frequency response at different frequencies, circles are shown that determine regions with a different maximum magnitude of the closed-loop response above unity gain. The smaller the circle that the graph will intersect, the higher this magnitude will be. These circles do not have their centre at the  $-1$  point on the real axis but on a shifting point to higher real numbers for a lower magnitude. These circles are derived from the magnitude of the complementary sensitivity function in the Nyquist plot for a certain level of  $Q$ :

$$|T(s)| = \frac{|GC|}{|1 + GC|} = \frac{\sqrt{\text{Re}^2 + \text{Im}^2}}{\sqrt{(1 + \text{Re})^2 + \text{Im}^2}} = Q \quad (4.29)$$

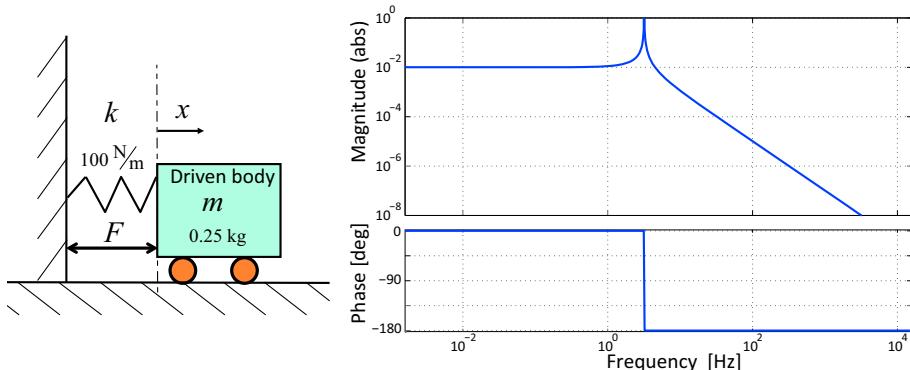
With a bit of algebra this results in:

$$\left(\text{Re} + \frac{Q^2}{Q^2 - 1}\right)^2 + \text{Im}^2 = \left(\frac{Q}{Q^2 - 1}\right)^2 \quad (4.30)$$

Which is a circle with its centre on the real axis at a distance  $D$  from the origin and with a diameter  $d$ :

$$D = -\frac{Q^2}{Q^2 - 1} \quad \text{and} \quad d = \frac{Q}{Q^2 - 1} \quad (4.31)$$

Modelling software like MATLAB automatically shows these circles and although unfortunately often with only the dB level mentioned, they are useful to get a first estimation of the expected closed-loop response of the system. In further examples of the Nyquist plot only the  $Q$  level will be mentioned to adhere to the preferred working with non-logarithmic numbers.



**Figure 4.21:** The undamped mass-spring system and the Bode-plot of its transfer function that is used as example to illustrate PID-control with some real data.

#### 4.3.4 PID-control of a mass-spring system

The previous example of the CD-player introduced P-control and D-control to create a virtual spring and damper. To create a real PID-controller another control action is added with its own beneficial effect, integrating control or *I-control*. According to control theory a PID-controller is defined by the following relation between the input error  $e$  and the output  $u$  of the controller in the time and frequency domain:

$$\begin{aligned} u(t) &= k_p e(t) + k_i \int_0^t e(\tau) d\tau + k_d \frac{de(t)}{dt} \\ u(s) &= e(s) \left( k_p + \frac{k_i}{s} + k_d s \right) \Rightarrow \\ C_{pid}(s) &= \frac{u(s)}{e(s)} = \left( k_p + \frac{k_i}{s} + k_d s \right) \end{aligned} \quad (4.32)$$

To show the effect of PID-control with practical values a PID-controller is designed for a simple mass-spring system in a comparable setting as the previous example of the PD-control of a CD player. In this case the body is connected to the stationary reference with a very compliant spring with a spring-constant  $k = 100 \text{ N/m}$  in order to show the difference with a pure inertial mass situation like with the CD player. The mass  $m$  of the body equals  $0.25 \text{ kg}$  and very little passive damping is present in the system ( $c \approx 0$ ) which means that the mechanical damping is neglected for this analysis. With these values and the gain of amplifier, actuator and sensor equal to

one, the transfer function of the plant becomes as follows:

$$G(s) = \frac{1}{ms^2 + k} \quad (4.33)$$

$$= \frac{1}{0.25s^2 + 100}. \quad (4.34)$$

At low frequencies the spring-line  $1/k = 0.01$  defines the transfer function. At  $\omega_0 = \sqrt{k/m} = 20$  rad/s the system has a resonance with an infinite high peak due to the absence of damping. At  $\omega \gg \omega_0$  the transfer function is dominated by the mass-line ( $1/ms^2$ ) with a  $-2$  slope and a phase lag of  $180^\circ$ .

For this system the PID feedback controller will be designed with the following control objectives:

- The control bandwidth as defined by the unity-gain cross-over frequency is 100 Hz.
- The feedback controlled system has to be *asymptotically stable*, which means that all the poles are located in the left-half of the Laplace plane.
- Sufficient damping of the feedback controlled system which implies no or little oscillations.
- The feedback controlled system must have zero steady-state error in response to the reference signal as well as disturbing forces acting on the controlled position of the mass.

The desired unity-gain cross-over frequency of 100 Hz is equivalent to  $\omega_c = 628$  rad/s, which is at a higher frequency than the resonance frequency of the mass-spring oscillator. In that region the frequency response of the plant is dominated by the mass-line ( $1/ms^2$ ). For this frequency the magnitude of the transfer function of the plant  $G(s) \approx 1 \cdot 10^{-5}$ .

#### 4.3.4.1 P-control

With a pure P-control gain of  $k_p = 1 \cdot 10^5$  the unity-gain cross-over frequency  $\omega_c$  of the transfer function of the plant is raised to the required value of 628 rad/s (100 Hz). For stability the total phase lag of the open-loop system has to be smaller than  $180^\circ$  at the unity-gain cross-over frequency. For the system discussed here a pure P-control gain results in a marginally stable system that shows in closed-loop a sharp resonance peak at the cross-over frequency with a sharply changing phase shift from  $0^\circ$  to  $-180^\circ$ . To add stability

and robustness to the feedback controlled system a differentiating action (D-control) is added that generates a phase lead around the unity-gain cross-over region, adding damping to the oscillatory system.

**Necessary reduction of the P-control gain due to D-control:** To add sufficient phase lead (and damping) the differentiating action (D-control) should start according to the “rule of thumb” at  $0.33\omega_c$ . This means that above this frequency the system shows a  $-1$  slope instead of the original  $-2$  slope of the mass-line. As was shown at the CD player example this requires that the gain  $k_p$  is reduced by this factor three in order to retain a loop-gain of one at 628 rad/s. As a result  $k_p$  is chosen to be equal to  $3.3 \cdot 10^4$ .

#### 4.3.4.2 D-control

The differentiating action (D-control) is used to add phase lead and damping to the system around the unity-gain cross-over frequency  $\omega_c$ . The differentiating term of the PD-controller ( $k_p + sk_d$ ) starts to dominate at  $\omega_c/3$ . Rewriting this PD-controller as

$$C_{pd}(s) = k_p \left(1 + s \frac{k_d}{k_p}\right), \quad (4.35)$$

shows that this happens at

$$\frac{k_p}{k_d} = s = \frac{\omega_c}{3}. \quad (4.36)$$

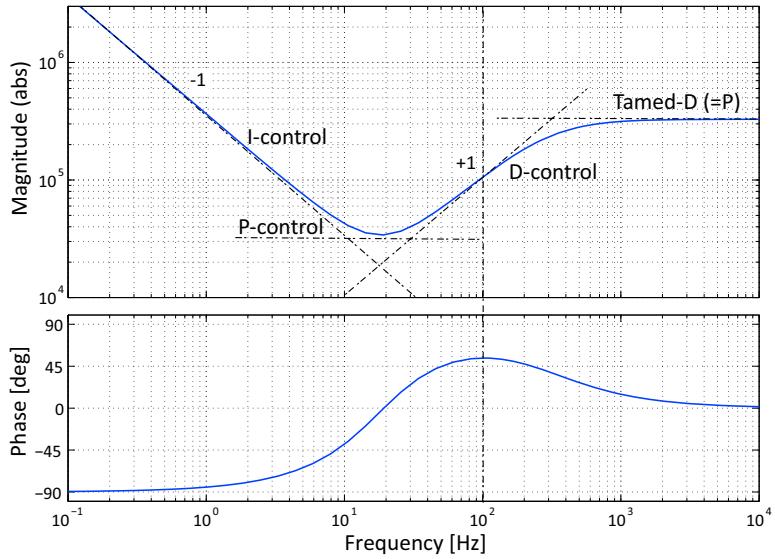
This means that

$$k_d = 3 \cdot \frac{k_p}{\omega_c} = \frac{3 \cdot 3.3 \cdot 10^4}{628} \approx 160. \quad (4.37)$$

The differentiating action is “tamed”, which means that it is limited to only the area around 100Hz, in order to provide a steeper roll-off of the controller at high frequencies and to limit the control effort at those frequencies. According to the “rule of thumb” the frequency above which the differentiation is terminated should be a factor of 3.3 above the cross-over frequency. Therefore a pole is added to the D-control gain at  $\omega_t = 3.3 \cdot \omega_c \approx 2000$  rad/s.

As a result the transfer function of the D-control action becomes:

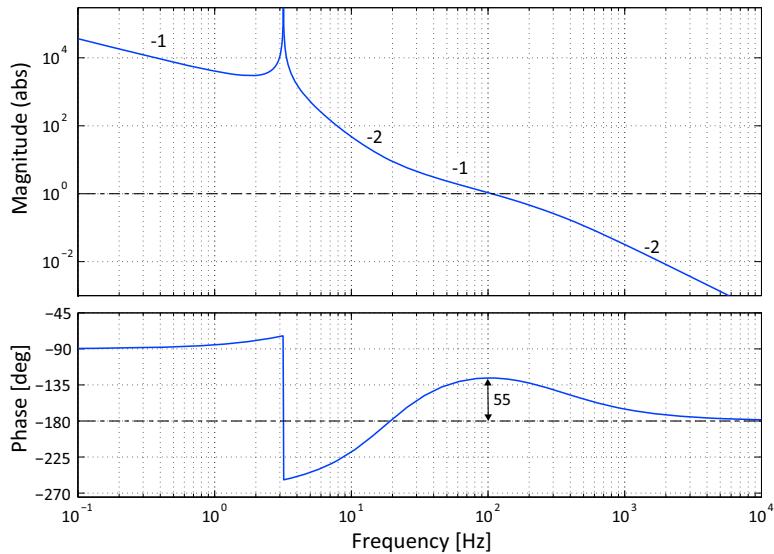
$$k_d = \frac{160}{5 \cdot 10^{-4}s + 1} \quad (4.38)$$



**Figure 4.22:** Bode-plot of the PID controller. From left to right first the  $-1$  slope of the I-control line is followed until it intersects with the horizontal P-control line at 10 Hz. At the intersection of the P-control line with the  $+1$  slope of the D-control line at 33 Hz, the magnitude increases again until the intersection at 330Hz with a second horizontal line, located a factor 10 above the P-control line, determined by the ratio of the differentiating and “taming” frequency where the controller transfer function becomes proportional again. At 100Hz this controller has the required gain of  $1 \cdot 10^5$  with a phase lead of around  $55^\circ$ , giving an equal phase margin for robustness.

#### 4.3.4.3 I-control

An integrating action is added to the system in order to increase the loop-gain at low frequencies and achieve a zero steady-state error in response to the reference signal as well as in response to disturbances. In the frequency domain, integration equals the addition of an  $s$  term in the denominator of the transfer function and as a consequence it adds  $90^\circ$  of phase lag to the system. In order to not affect the beneficial phase lead of the D-control too much, the integrating action should stop another factor of 3.3 lower than the starting point of the D-control action, so at  $\omega_i \approx 0.1 \cdot \omega_c \approx 60$  rad/s. At this angular frequency the term  $k_i/s$  has to become equal, and at higher frequencies eventually smaller, than  $k_p$ .



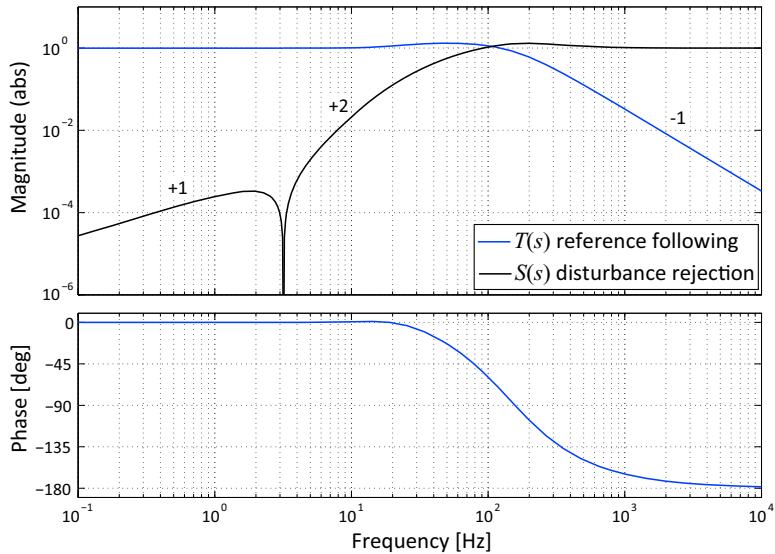
**Figure 4.23:** Bode-plot of the open-loop transfer function of the total system with the PID controller in series with the mass-spring system. Even though the eigendynamics of the mechanical system show a clear resonance at around 3 Hz the phase lag at the 100 Hz unity-gain cross-over frequency remains well below 180°.

Therefore:

$$k_i = k_p \cdot 60 \approx 1.8 \cdot 10^6 \quad (4.39)$$

These controller gains for  $k_p$ ,  $k_d$ , and  $k_i$  result in the Bode-plot of the transfer function of the controller shown in in Figure 4.22. The integrating action up to 10 Hz and the differentiating action from 33 Hz until 330 Hz with the corresponding phase behaviour are clearly visible.

Figure 4.23 shows the Bode-plot of the open PID-control loop, consisting of the PID controller in series with the mass-spring system, as defined in Equation 4.33. At low frequencies the integrating action is providing a -1 slope for a high gain. At 100 Hz the gain of the open-loop system is equal to one as required. Beyond the natural frequency of the uncontrolled mass-spring system at  $\approx 3.3$  Hz the system shows after the sharp peak at the resonance a -2 slope which is reduced to a -1 slope by the differentiating action around the unity gain cross-over frequency. The phase margin of this system is about 55°, being the distance between the phase lag of the system ( $\approx 125^\circ$ ) at the cross-over frequency of 100 Hz and the maximum limit of 180° phase lag. The amplitude margin of this system is infinite because in



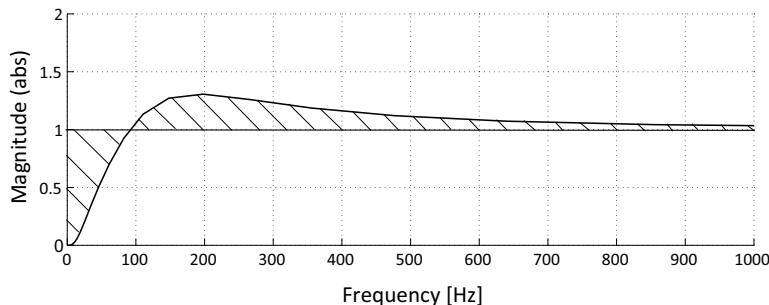
**Figure 4.24:** Bode-plot of the closed-loop PID feedback controlled mass-spring system. Both the complementary sensitivity function  $T(s)$ , representing the capability to follow the reference as the sensitivity function  $S(s)$ , representing the disturbance-rejection capability, are shown. The original resonance has disappeared and is turned into a strong rejection of disturbances at that frequency.

this ideally modelled example the open-loop system reaches the  $180^\circ$  phase lag asymptotically only at infinitely high frequencies.

Figure 4.24 shows the Bode-plot of the closed-loop system. The phase plot is generally not meaningful anymore in the closed-loop frequency response as it does not give any further information on stability. For precision systems that have to move synchronously, like with the wafer scanner of Chapter 9, the closed-loop phase gives information on the time delay of the movement and in that case it is still important information.

In this case the phase plot just shows the standard shape of a well damped mass-spring system with a natural frequency at 100 Hz, corresponding with the blue magnitude line of the complementary sensitivity function  $T(s)$ . At low frequencies the closed-loop gain is equal to one which means that the system follows the reference signal with only a very small error. Beyond the control bandwidth the feedback controlled system rolls off to higher frequencies with a  $-2$  slope. The fact that the feedback controlled system is well damped can be observed by the absence of a resonance peak.

As mentioned with the previous example of the CD player, the roll off beyond



**Figure 4.25:** The sensitivity function  $S(s)$  of the PID-feedback controlled mass-spring system plotted in a linear scale. This representation clearly shows the increased sensitivity of this actively controlled system at frequencies above the bandwidth of 100Hz. When the shaded area below the magnitude of one is increased for better rejection of low frequencies, the shaded area above a magnitude of one also increases. This *waterbed effect* is inherent to feedback systems and can not be avoided.

the control bandwidth with a  $-2$  slope is because the differentiating action is “tamed”. If that were not done, only a damped transmissibility behaviour would be realised, with a  $-1$  slope attenuation and an infinitely high control gain at high frequencies. For this reason it is emphasised again that this additional pole to stop the differentiator is important.

The sensitivity function  $S(s)$  shows small values and a  $+1$  slope at low frequencies. This corresponds with a good disturbance-rejection due to the integrator that acts up to 10 Hz. At the natural frequency of the mass-spring system a sharp dip is observed in  $S(s)$  with a steeper slope at higher frequencies, which is due to the high gain by the resonance. It is very valuable to be aware of this effect as this high gain can help to suppress external disturbances at the resonance frequency of the mechanical system. This effect is very much counter-intuitive and can only be applied when a system suffers from a continuous disturbance at a fixed frequency. Examples are not ideally rotating systems like the eccentricity in the track of a CD-player or a hard-disk drive and the forces by mass unbalance in high speed spindles of machining centres.

Beyond the unity-gain cross-over frequency  $S(s)$  is settling around a value of one after a frequency range where the magnitude is larger than one, due to the so called *waterbed effect*. This waterbed effect has got its name from the property of any feedback control system that an increase in disturbance-rejection at low frequencies, due to different controller settings, automati-

cally causes an increase of the sensitivity at frequencies above the bandwidth. This is indicated by the *Bode Sensitivity Integral*:

$$\int_0^\infty \ln|S(\omega)| d\omega \quad (4.40)$$

This integral is zero when the open-loop transfer function  $L(s)$  of the dynamic system has at least two more poles than zeros and when all poles are located in the left half of the Laplace plane. This condition is always true with stable feedback controlled positioning systems.

The reason why the waterbed effect is not so clearly visible in a normal Bode-plot is due to its logarithmic scales. Figure 4.25 shows the magnitude of  $S(s)$  on a double linear scale. With this representation the frequency range above 100 Hz with an increased sensitivity appears to be much larger than the frequency range below 100 Hz, while in a normal Bode-plot this difference is not as pronounced. Also the magnitude is relatively reduced with double logarithmic scales. In the linear representation the benefit of feedback seems to have vanished but linear scales are also not representative for the real value of feedback in the low-frequency area. A reduction of a factor 100 or a factor 1000 of the disturbances at low frequencies can not be distinguished on a linear scale. In most motion control systems the disturbances occur more in the low-frequency area where the process disturbance  $d(s)$  has typically a sort of  $1/f$  spectrum over the frequency range. This means that in practice the beneficial effect of feedback in the low-frequency region by far outweighs the disturbance amplification at higher frequencies due to the waterbed effect. It is also for this reason that logarithmic scales are preferred as long as the mechatronic designer is aware of the negative effect on disturbance rejection just above the unity-gain cross-over frequency for those cases that significant disturbances are also present in that frequency region.

### 4.3.5 PID-control of more complex systems

The previous sections presented rather simplified systems that in reality hardly ever exist. In this section two examples will be presented to give a flavour of the variations that can be encountered when designing controlled motion systems. The first example deals with the control of a system with negative stiffness like a magnetic bearing. The second example explains a method how to deal with the resonances due to higher order eigenmodes that are always present in any mechanical structure.

#### 4.3.5.1 PID-control of a magnetic bearing

In Section 4.1.1 it was mentioned that a system with a negative stiffness gives a pole in the right-half of the Laplace plane. It was also mentioned that PID-control can be used to shift the pole to the stable left-half of the Laplace plane by adding a positive stiffness and damping. Controlling of such a system is important as some of the electromagnetic actuators that are described in the next chapter show a negative stiffness. Especially magnetic bearings are an example of such a system and as a consequence they can only work with active feedback control. To illustrate how this control is achieved Figure 4.26 shows the Bode-plot from the following transfer function of a full plant in one direction (SISO), consisting of an actuator with a gain equal to one and a (negative!) stiffness  $k_n = -10^4 \text{ N/m}$ , a moving mass  $m_m = 0.1 \text{ kg}$  and a position sensor also with a gain of one:

$$C_t(s) = \frac{x}{F} = \frac{1}{m_m s^2 + k_n} = \frac{1}{0.1 s^2 - 10^4} \quad (4.41)$$

$$C_t(\omega) = \frac{1}{-0.1\omega^2 - 10^4} = -\frac{1}{0.1\omega^2 + 10^4} \quad (4.42)$$

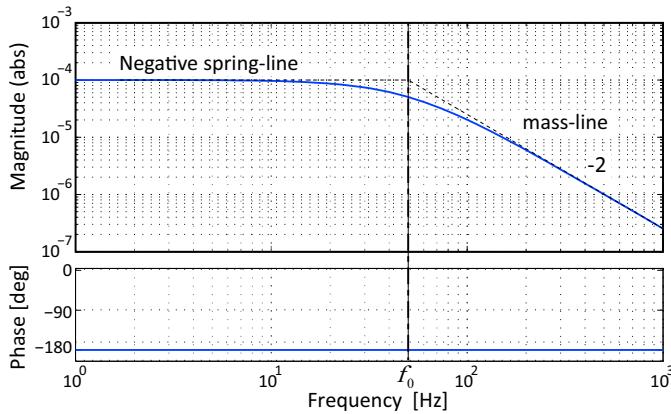
At first sight the magnitude plot looks like a well damped mass-spring system but the phase plot is fully different with  $-180^\circ$  phase over the entire frequency band, caused by the negative stiffness value. This negative sign implies that the position is always in the opposite direction of the force.

With increasing frequency the compliance of the mass-line will take over the total compliance from the negative spring-line at their intersection frequency<sup>4</sup>  $f_0$ :

$$f_0 = \frac{1}{2\pi} \omega_0 = \frac{1}{2\pi} \sqrt{\frac{k_n}{m_m}} = \frac{1}{2\pi} \sqrt{\frac{10^4}{0.1}} \approx 50 \quad [\text{Hz}] \quad (4.43)$$

---

<sup>4</sup>This intersection frequency is not named "natural-" or "eigenfrequency" because no resonance occurs and no eigenmode is observed.

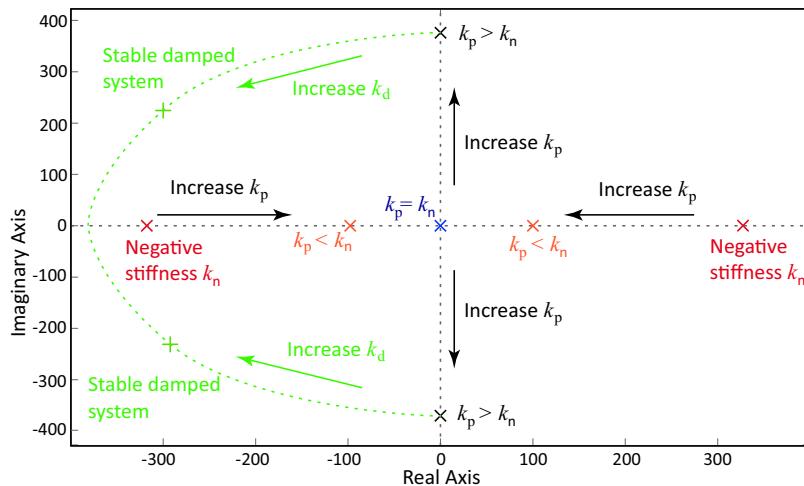


**Figure 4.26:** The Bode-plot of the transfer function of a magnetic bearing shows a frequency response without resonance and with a constant phase of  $-180^\circ$  due to the negative stiffness value.

The phase remains constant over the entire frequency band because the mass-line also has a phase of  $-180^\circ$ . Next to the negative sign a second observed difference with a regular mass-spring system is the lack of resonance. This is caused by the fact that the denominator does not become zero at any value of  $\omega$  and this corresponds again with the location of both poles on the real axis at a value of  $\pm\sqrt{k/m} = \pm\sqrt{10^5} \approx 320$ . With this information the design of a PID controller can be realised according to the same steps made in the previous part of this section. Still some things will show to be quite different because of the negative stiffness.

First P-control is applied in a magnitude that is at least larger than the magnitude of the spring-line. In mechanical terms this means that after closing the negative feedback loop the positive stiffness by the P-control action will be larger than the negative stiffness of the actuator. For reasons of robustness against variations in the negative stiffness the positive control stiffness  $k_p$  should be chosen minimally as large as the maximum value of the negative stiffness that can occur. A practical value is at least a factor two larger than the average value of the negative stiffness. This choice results in a total positive stiffness that is always equal to or larger than the magnitude of the negative stiffness.

The effect on the poles is illustrated in Figure 4.27 and it shows that with only P-control, the closed-loop response would become identical to a regular undamped mass-spring system. The equivalent natural frequency of this system would correspond with the unity-gain cross-over frequency  $f_c$  of the combined negative-stiffness actuator with the P-control gain  $k_p$  at a



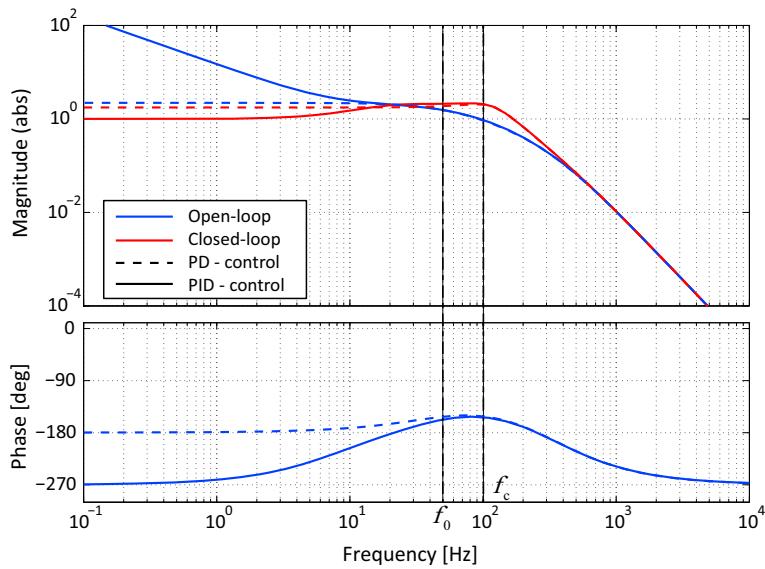
**Figure 4.27:** The poles of the unstable magnetic bearing system with a moving mass  $m_m = 0.1 \text{ kg}$  and a (negative!) stiffness of  $k_n = -10^4 \text{ N/m}$  are located at  $\pm 320$  on the real axis. The poles are brought to the left-half of the Laplace plane in two steps. First P-control with an increasing value of  $k_p$  is added until at  $k_p = 2.3 \cdot 10^4$  an undamped system is created with both conjugate complex poles at  $\pm 360$  on the imaginary axis. In the second step D-control is added to create damping by shifting the poles over a circle to the left. Note that the circle is an ellipse because of the difference in scales.

frequency that is equal to or slightly larger (square root of stiffness!) than  $f_0$ .

The second step is the addition of D-control according to the rules of thumb for D-control. In this case the proportional gain should however not be reduced because then the total loop gain at the start of the differentiating action would become lower than one with a phase of more than  $-180^\circ$ , leading to instability. D-control will shift the poles to the left over a circle that has been defined for a damped mass-spring system in Equation (3.48) of Chapter 3. As a further consequence the D-control action will increase the unity-gain cross-over frequency and in practice the open-loop unity-gain cross-over frequency will become approximately twice the value of the intersection frequency  $f_0$  of the mass- and the spring-line of the magnetic bearing.

This effect leads to an important conclusion regarding magnetic bearings:

The negative stiffness of a magnetic bearing **sets a lower limit** to the bandwidth that has to be realised in the closed-loop system.

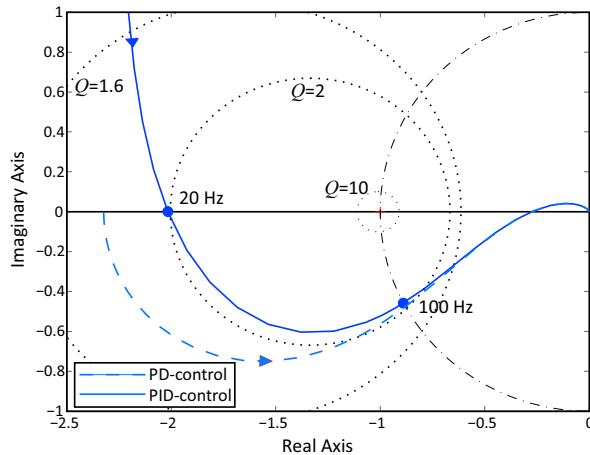


**Figure 4.28:** The open-loop and closed-loop Bode-plot of the magnetic bearing with negative stiffness with PD-control (dashed lines) and PID-control (solid lines). In closed-loop the frequency response shows a magnitude larger than one  $Q = 2$  that stretches over a wide frequency range until  $f_c$ . With PD-control starting at 0 Hz and with PID-control starting around 20 Hz.

When other dynamic problems like resonating mode shapes and delays in the electronics prevent the realisation of such a high bandwidth the magnetic bearing can not be made stable.

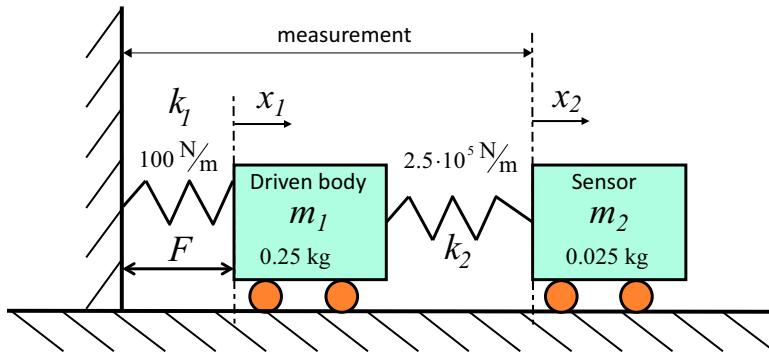
This statement is in principle valid for any system that is unstable in open-loop, like an inverted pendulum or a rocket.

As a last addition I-control is applied in order to increase the rather limited open-loop gain below  $\omega_i = 0.1\omega_c$ . Even though it may seem that a significantly higher proportional gain than the negative stiffness value would create a very stiff system, this higher proportional gain only creates a parallel stiffness **after closing the loop**. In the open-loop situation, the multiplication results in a loop gain that is equal to the **number of times** that the proportional gain is chosen larger than the negative stiffness, which is often not sufficient for disturbance suppression at lower frequencies. Furthermore a very high  $k_p$  far above the negative stiffness would give a very high unity-gain cross-over frequency with an increased risk on instability to high-frequency resonances by eigenmodes of the mechanical system.



**Figure 4.29:** The Nyquist plot of the open-loop frequency response of the magnetic bearing with PD- and PID-control explaining the magnitude of the closed-loop frequency responses by the position relative to the  $Q$ -circles.

Figure 4.28 shows the open-loop and closed-loop frequency response of the combined PD- and PID-controller with the plant with negative stiffness. In this case  $k_p = 2.3 \cdot 10^4$ , which is 2.3 times the value of the negative stiffness. In feedback this value adds to the negative stiffness to a total value of  $k_t = 1.3 \cdot 10^4$  and the poles of the closed-loop system have an imaginary value of  $\pm\sqrt{k_t/m} = \pm\sqrt{1.3 \cdot 10^5} \approx 360$ , as shown in Figure 4.27. The D-control action goes from 30 Hz until taming at 300 Hz giving a unity-gain cross-over frequency of 100 Hz. I-control works below 10 Hz. The closed-loop bandwidth is 100 Hz and the resonance peak is less damped ( $Q = 2$ ) than with the previous system with a positive stiffness value as the phase lead is only around 30°. With PD-control this effect is even more prominent as it stretches until 0 Hz. This effect is fully caused by the  $-180^\circ$  of the negative stiffness and can be better analysed with the Nyquist plot of Figure 4.29. With PD-control the magnitude starts at  $k_p \times k_n = 2.3$  and runs around the  $Q = 2$  circle. Over the frequency band ranging from 0 Hz until just more than 100 Hz the graph remains between the  $Q = 1.6$  and the  $Q = 2$  circle which means that in closed-loop the magnitude will be around 1.8 in that frequency range, which is observed in the Bode-plot. With PID-control the graph runs just inside the  $Q = 2$  circle from 20 Hz to 100 Hz, which results in the observed magnitude of the closed-loop frequency response in that frequency range in the Bode-plot, while the closed-loop response below 20 Hz



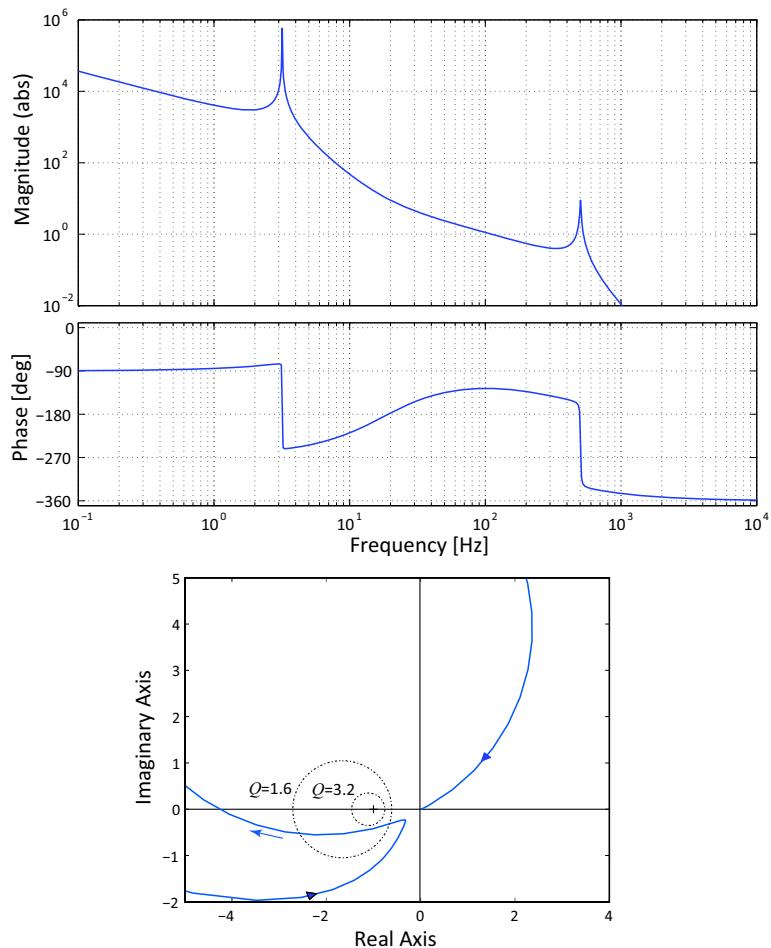
**Figure 4.30:** When the sensor is connected to the driven body with a compliant mounting, the mass of the sensor will decouple at high frequencies due to the second eigenmode, where the two bodies vibrate in opposite directions by the elastic coupling of  $k_2$ .

approaches a value of one because of the high loop gain due to the integrator. A reduction of the proportional gain below the example value  $k_p = 2.5 \cdot 10^4$  will cause the graph on the Nyquist plot to approach the  $-1$  point on the real axis, resulting in a higher  $Q$  over a wider range of frequencies until at-half the presently chosen value ( $k_p = 1.15 \cdot 10^4$ ), the system will start to resonate at 20 Hz. Optimal tuning of such a magnetic bearing is clearly not trivial as the optimisation depends on many parameters, like the expected frequency spectrum of the disturbances and the frequency range of operation.

#### 4.3.5.2 Eigenmodes above the desired bandwidth

At the presentation of D-control it was demonstrated that the D-action should be terminated (tamed) above a frequency of approximately three times the cross-over frequency. Next to the fact, that differentiation until infinite frequencies is impossible anyway, this measure would be beneficial in case of eigenfrequencies of the system at higher frequency levels. In general it is indeed a good starting point, when designing the feedback system, to try to keep the magnitude of the frequency response at these higher eigenfrequencies below one, in order to guarantee stability.

In the following, this requirement will be examined in more detail and it will be shown that it is possible to reduce this requirement to some extent by manipulating the phase. For the example, the same mass-spring system of the previous section is used, but now the position sensor is more real. It is not mass-less co-located directly on the primary body, but it consists of



**Figure 4.31:** Bode and Nyquist plot of the open-loop transfer function  $L(s)$  of the PID-controlled mass-spring system when the sensor is mounted on a body that decouples dynamically at 500Hz with  $Q \approx 20$ . The graph in the Nyquist plot passes the  $-1$  point on the real axis at the right hand side and instability will occur after closing the loop. It illustrates the necessity to keep the amplitude of the resonance below one in order to maintain stability.

a body with a mass of  $25 \cdot 10^{-3}$  kg with a mounting stiffness of  $2.5 \cdot 10^5$  N/m, as shown in Figure 4.30. In Section 3.3.1 of Chapter 3 on the dynamics of multiple coupled bodies the configuration appeared to result in a second eigenmode with an eigenfrequency of  $\approx 500$  Hz. Because in practice such situations show a very limited damping, a  $Q$  of  $\approx 20$  is used in the modelling.

Figure 4.31 shows both the Bode-plot and the Nyquist plot of this situation. The Nyquist plot is especially useful in this situation as it shows in clear detail, what happens around the  $-1$  point on the real axis with respect to closed-loop stability in this more complex dynamic feedback system. The phase at 500 Hz due to the second eigenmode changes from almost  $-150^\circ$ , the mass-line combined with the D-control action, to below  $-180^\circ$ . This phase shift occurs in the frequency range where the magnitude becomes larger than one due to the resonance peak of the second eigenmode and as a consequence instability will occur.

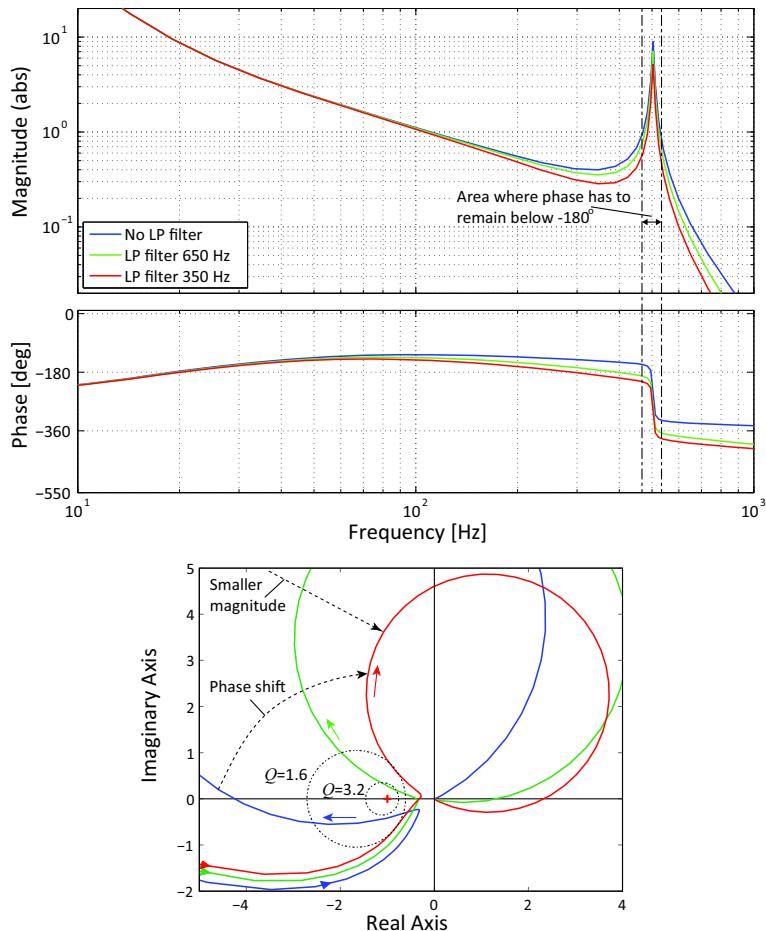
The Nyquist plot clearly indicates that the  $-1$  point on the real axis is kept at the right hand side when passing. Only when the magnitude of the second eigenmode is kept below one by a reduction of the loop gain this instability condition would not occur, but then the resulting bandwidth of the controlled system would be reduced to below  $\approx 20$  Hz.

### Shifting the phase

Different approaches are possible to overcome this problem. Often a notch filter is applied, with an inverse characteristic of the resonating eigenmode to suppress this specific resonance. An example was shown with the model-based feedforward controller of Section 4.2.1. This method requires however an almost perfect tuning of the frequency of the filter to the resonance, particularly with a very high  $Q$ -factor and in mass production every controller will then be different due to the spread in the system dynamics by production tolerances.

One might also consider adding damping to the decoupling spring  $k_2$  although generally this spring is not compliant on purpose but the result of a not infinite stiff mounting. Adding damping to a mounting is very hard to do and sometimes even results in a reduced stiffness, shifting the second mode to lower frequencies, like with rubber mounts.

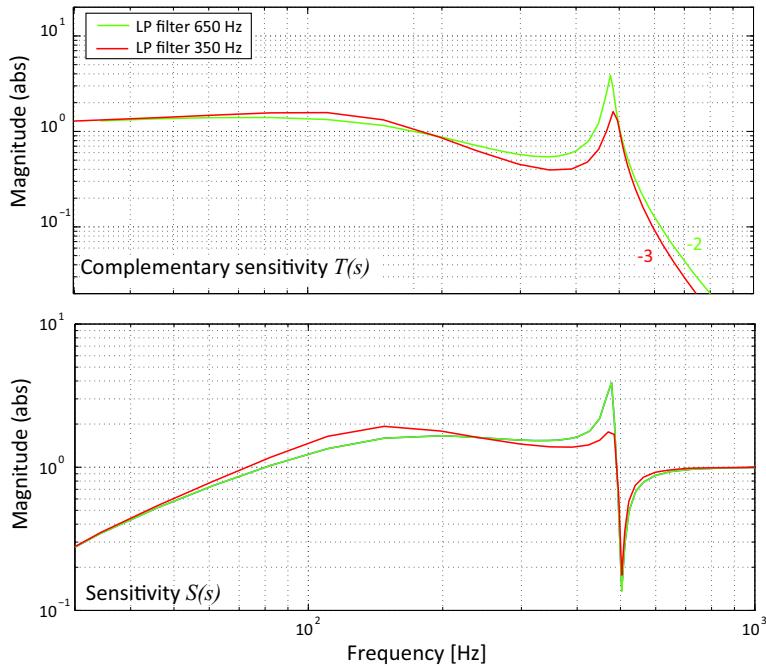
Another method, more related to active control, is based on the understanding that the problem occurs because the phase is almost  $-180^\circ$ . When the phase lag would be larger, a high magnitude above one would no longer give a problem as long as the  $-1$  point on the real axis is passed at the left hand side. For the given example this can simply be arranged by inserting a simple low-pass filter in the loop at a frequency around the eigenfrequency of the decoupling mass. Combined with the first-order low-pass filter used for the “taming” of the differentiator, this second low-pass filter forms a second-order low-pass filter. This combination can also be replaced by a less damped second-order low-pass filter in order to limit the phase shift at the



**Figure 4.32:** Inserting an additional low-pass filter in a PID-controlled mass positioning system will only slightly decrease the phase margin, but changes the phase at the decoupling mass resonance to below  $-180^\circ$ . The Nyquist plot shows more clearly the beneficial effect. When filtered at 350 Hz the  $Q = 1.6$  circle is just touched.

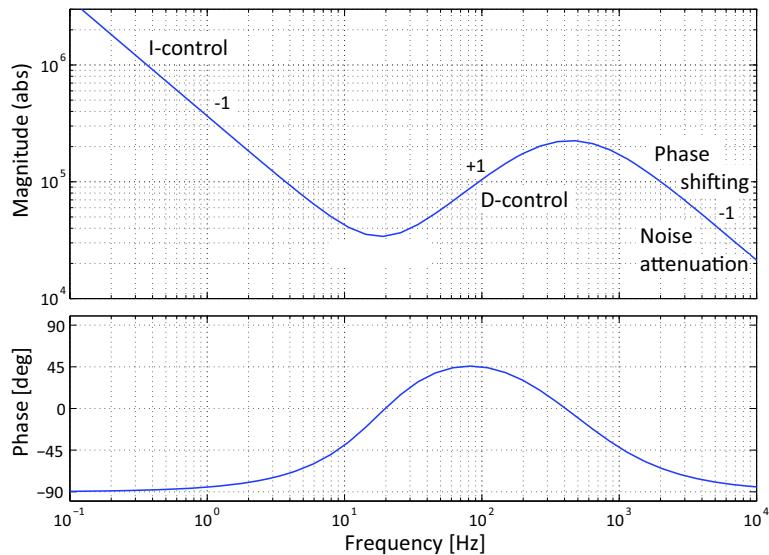
unity-gain cross-over frequency, but this is a detail that can be adapted when a further optimisation of the system is required. For the understanding of the principle it is useful to treat them first as two separate filters.

Figure 4.32 shows the effect of the additional low-pass filter on the open-loop transfer function of the example PID-controlled mass-spring system. Next to the situation without the additional low-pass filter, also the open-loop response of the system with a low-pass filter at 350 or 650 Hz is shown. The



**Figure 4.33:** The closed-loop Bode-plot of the sensitivity functions of the PID-controlled mass positioning system with decoupling mass show the effect of the low pass filter frequency on the magnitude of these functions at the cross-over frequency and the eigenfrequency of the decoupling mass. At 350 Hz these magnitudes are equally balanced. This optimal setting is also confirmed by the sensitivity plot at the lower side.

Bode-plot clearly demonstrates that such a filter has only a limited negative effect on the phase margin at the cross-over frequency of 100 Hz, while the phase at the resonance has definitely come below  $-180^\circ$ . In this more complex PID-control situation the Nyquist plot is an indispensable tool as it more clearly shows the effects regarding stability. With the additional first-order low-pass filter, stability is achieved as the  $-1$  point on the real axis of the Nyquist plot is passed at the left hand side, while without a low-pass filter the system would be unstable. An optimal situation is created when the filter frequency is at 350 Hz resulting in a symmetrical left hand passage of the  $Q = 1.6$  circle, indicating an equal magnitude of the closed-loop response for the two frequencies that are closest to the circle. It results in a maximum amplitude rise of 60 % both at the cross-over frequency and at the second eigenmode. This optimum is confirmed in the complementary sensitivity  $T(s)$  and sensitivity  $S(s)$  Bode-plot of Figure 4.33.



**Figure 4.34:** The PID-control transfer function with additional phase shifting low pass filter can be seen as an optimal motion controller of a mechanical system.

A clear benefit of this method is that the influence due to the resonance of a second eigenmode with an amplitude larger than one is effectively suppressed by the feedback. The original  $Q$  value from the uncontrolled system is reduced from more than twenty to less than a factor two by tamed PID-control with an additional first order low-pass filter. Another advantage of the additional low-pass filter is the  $-3$  slope of the sensitivity function  $T(s)$  at higher frequencies which means that the resonances of higher eigenmodes and the high-frequency sensor noise will be even better suppressed.

#### 4.3.5.3 “Optimal” PID control

The described PID-control transfer function with a combination of the three basic PID-control gains and a second-order low-pass filter for taming the D-control action and shifting the phase at the first resonance is shown in Figure 4.34. In practice this configuration has become a standard in low-stiffness mass-positioning feedback systems with force actuation by Lorentz actuators. It has proven to be an optimal choice in that application area with respect to robustness against system variations and tolerances.

In control theory, the term *optimal control* generally has another though closely related meaning, where a mathematical optimisation method is

used to find a control algorithm for a given system, such that a certain optimisation criterion is achieved. Optimality of this PID-control scheme with second-order low-pass filter is understood as the best combination of robustness and performance so corresponding with the “practicality” criterion of dynamic positioning systems.

#### 4.3.5.4 Open-loop and closed-loop

This section has clearly demonstrated the creation of a virtual spring and damper with an additional integrating function by PID-feedback control. It is unfortunate that a mechanical equivalent for the integral gain  $k_i$  is not readily available. The integrating action provides a high controller gain  $C$  at low frequencies. For that reason the closed-loop frequency response becomes equal to one in that low-frequency region, independent of the transfer function of the plant. For that reason the integral action is acting like a “super-spring” that increases the loop-gain at low frequencies and gradually reduces its distortion to zero. As a result the steady-state error of the feedback controlled system in response to a constant reference signal also becomes equal to zero.

These virtual elements are virtually placed in parallel to the physical spring and damper of the plant and the combination acts as if the virtual and real stiffness values and damping coefficients are simply **added**. It is very important however to realise that this effect only occurs **upon closing the loop**. In the open-loop situation, when the output  $y$  is no longer connected to the summing point with the reference signal  $r$ , the controller is placed in series before the plant and in that case the transfer functions of the controller and the plant are **multiplied**. This difference often leads to confusion as can be illustrated by examining only the proportional branch of the system as example, consisting of the stiffness of the spring and the proportional gain.

From the previous chapter it is known that the transfer function of a physical spring with stiffness  $k$  acts open-loop as a compliance with a value of  $k^{-1}$ . In the open-loop situation the total compliance of this transfer function can be increased with P-control with a value needed to bring the total compliance to unity gain ( $\Rightarrow k_p = k$ ). After closing the loop this value of  $k_p = k$  would result in an additional spring with a value of  $k$  giving a total stiffness of  $2k$  in combination with the physical spring. At first sight there appears no relation whatsoever between the unity gain open-loop value and the double stiffness in closed-loop but when calculating the closed-loop frequency response, the open-loop transfer function of  $G(s)C(s) = 1$  results in

a feedback complementary sensitivity function  $T(s) = GC/1 + GC = 0.5$ . This value indicates that the closed-loop compliance equals half the open-loop compliance and implies that the stiffness of the closed-loop system has become twice the stiffness of the physical spring, which corresponds with the double stiffness that was determined by the addition.

This example shows that it is necessary to be always aware that the virtual elements created by the feedback loop are only present when the feedback loop is closed. In the open-loop situation these element-terms are meaningless. The open-loop analysis is used for tuning the controller regarding stability and loop gain as explained in this section and should not be mixed with these element-terms.

## 4.4 State-space control representation

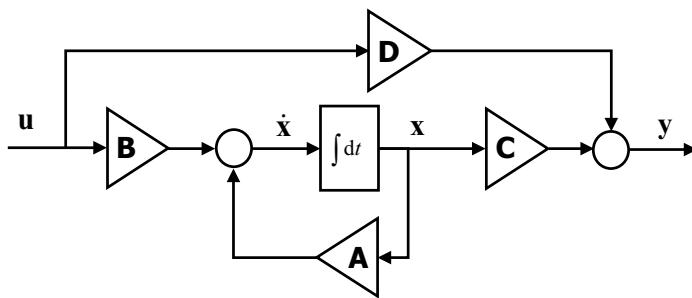
Generally a single body positioning system needs to be positioned in six orthogonal coordinate directions necessitating six actuators and six sensors. In control system terms this positioning system is called a *Multiple Input Multiple Output* system (MIMO) and complications can occur due to the mutual interference or *cross-coupling* between each direction. When all sensors and actuators are well aligned in six unique directions, the system can be reduced to a combination of six independent SISO subsystems and even when they are not aligned a SISO approach can often be achieved by means of transformation matrices under the condition of a sufficiently deterministic behaviour. An example of this principle will be presented in Chapter 9 with the control of a wafer stage. In situations, when there are strong (non-linear) cross-coupling terms between the several inputs and outputs that can not be compensated in a straightforward deterministic way, this approach is not sufficient especially when higher frequency eigenmodes with a multitude of bodies also create cross-coupling terms.

The *state-space* representation has been introduced in control practice as a time-domain related method to be used in digital controllers to calculate the required MIMO control actions. The method is highly efficient because it breaks down an  $n^{\text{th}}$ -order differential equation into  $n$  first-order differential equations that are written in a matrix notation. With the state-space model, every successive control action can be calculated by straightforward vector-matrix operations.

The state-space model has three different variables, the inputs  $\mathbf{u}$ , the outputs  $\mathbf{y}$  and the *state-variables*  $\mathbf{x}$  that describe the momentary state of the system. Generally these variables are written as vectors because they contain a number of elements. The state-variable vector defines a *space* of which the axes are the state-variable elements in the vector. This has given the name state-space to this modelling method. The total number of state-variables, mostly just called “states”, is at least as large as the number of orders of the differential equation that is modelled.

The standard state-space notation is given by the following two equations that do not contain the  $(t)$  term as the state space representation is uniquely used in the time domain:

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} + \mathbf{Du}\end{aligned}\tag{4.44}$$



**Figure 4.35:** Graphical representation of a standard multi dimensional state-space system. The feedback section with matrix  $\mathbf{A}$  represents the eigendynamics of the system. The input section of the system consists for instance of the actuators with gain matrix  $\mathbf{B}$  and the output section of the system consists for instance of the sensors with gain matrix  $\mathbf{C}$ . The section with matrix  $\mathbf{D}$  represents the direct feed-through of the input to the output and is for instance caused by the cross-coupling from the actuator current to the sensor signal.

These equations show how the state-variable  $\mathbf{x}$  of a dynamic system and the system output  $\mathbf{y}$  evolve over time as a function of the matrices  $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ , the input  $\mathbf{u}$  and the initial conditions  $x_0$  of the state vector  $\mathbf{x}$ . The matrices are the eigendynamics matrix  $\mathbf{A}$ , the input matrix  $\mathbf{B}$ , the output matrix  $\mathbf{C}$  and a feed-through gain  $\mathbf{D}$ .

Figure 4.35 shows the structure of this standard state-space model in a graphical way. Like described at the figure, the different elements of these matrices all represent a certain physical property or function.

#### 4.4.1 State-space in relation to motion control

State-space motion control uses a state-space model of the full mathematical description of the plant, including actuators and sensors, where the controller actions are applied to the external connections of the uncontrolled system, the inputs and outputs. To illustrate what this means, a few examples will be shown that were introduced before with the analytical approach to motion control.

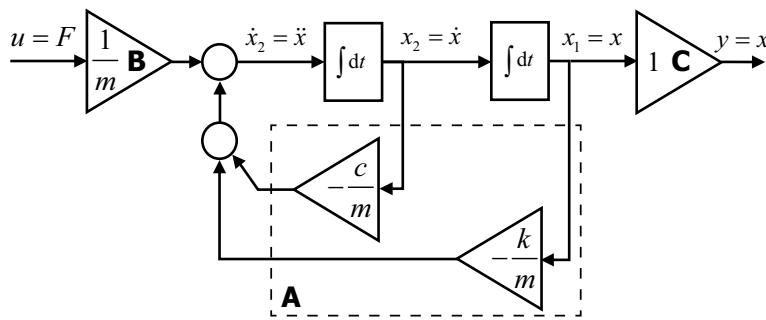
Chapter 3 presented a damped mass-spring system as a second-order system and when combined with one or more spring-coupled bodies, it became a higher-order system adding two orders for each coupled mass-spring system. In general the order  $n$  of a dynamic system is equal to the number of energy

containers in the system, like a spring ( $E = 0.5kx^2$ ) and a body ( $E = 0.5mx^2$ ). A damper does not contain energy. When the state-space model is used for controlling the plant, the order of the differential equations that describe the plant does not only include the mechanics but also the dynamics of the applied sensors, amplifiers and actuators. A first-order filter at the sensor introduces an additional state and the dynamics of the actuator also create corresponding states.

First the different matrices will be examined from the viewpoint of motion control:

- The eigendynamics matrix **A** includes the parameters that describe the dynamics of the uncontrolled system, including the dynamics of the power amplifier, the actuator, the mechanical system and the sensor. Its dimension is square ( $n \times n$ ).
- The input matrix **B** contains the simple gains or scaling factors of the actuator or power amplifier, but not their dynamics as these have to be included in matrix **A**, as stated above. The number of columns of the input matrix is equal to the number of inputs ( $i$ ) and the number of rows equals the number of states  $n$ .
- The output matrix **C** contains the simple gains or scaling factors of the measurement system, but also without the dynamic part as that also has to be included in matrix **A**, as stated above. The number of columns of the output matrix is equal to the number of states ( $n$ ) and the number of rows equals the number of outputs  $o$ .
- The feed-through matrix **D** is often not present in mechatronic system, but can for instance be caused by actuator-sensor cross-talk, like when using electromagnetic actuators and an Eddy-current sensor. Its dimension depends on the number of inputs and outputs with  $i$  columns and  $o$  rows.

These matrices and the states in the state-space model are illustrated with two examples, a damped mass-spring system, with and without a classical PID-controller.



**Figure 4.36:** Graphical representation of the scalar, single input and output (SISO), state-space representation of a mass-spring system with damping. The loops of the damper (velocity) and the spring (position) are shown separately.

#### 4.4.1.1 Damped mass-spring system

A simple single directional (SISO) damped mass-spring system with position  $x$  of the body is described in state-space as follows:

The state  $\mathbf{x}$  and its derivative  $\dot{\mathbf{x}}$  equal:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ \dot{x} \end{bmatrix} \quad \dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \ddot{x} \\ \ddot{\dot{x}} \end{bmatrix} \quad (4.45)$$

With this SISO example the input  $\mathbf{u}$  would be the force with a scalar value  $u = F$ . The output  $\mathbf{y}$  is equal to the position  $x$ , so it is also a scalar. The input matrix  $\mathbf{B}$  in this case is the conversion from force to acceleration with a mass factor and the eigendynamics matrix  $\mathbf{A}$  contains the damper and spring relations which is shown in the following expansion of the state-space equations:

$$\begin{aligned} \dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} &\implies \begin{bmatrix} \dot{x} \\ \ddot{x} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & \frac{-c}{m} \end{bmatrix} \cdot \begin{bmatrix} x \\ \dot{x} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} F \\ \mathbf{y} = \mathbf{Cx} + \mathbf{Du} &= [1 \ 0] \cdot \begin{bmatrix} x \\ \dot{x} \end{bmatrix} + [0]u = x \end{aligned} \quad (4.46)$$

When this is written out for the acceleration the following equation is obtained:

$$\ddot{x} = \frac{-k}{m}x - \frac{c}{m}\dot{x} + \frac{F}{m} \implies F = m\ddot{x} + c\dot{x} + kx \quad (4.47)$$

which is the standard force equation that was used in Chapter 3 to derive the frequency response of the system. Also with more elaborate systems

the transfer function is derived by applying the Laplace transform on the state-space model according to this example.

The state-space representation of this SISO mass-spring system is shown in a graphical way in Figure 4.36 where  $x_1 = x$  is the position,  $x_2 = \dot{x}$  is the velocity and  $\dot{x}_2 = \ddot{x}$  is the acceleration of the system. The feedback loops with the stiffness and damping terms will show to be very illustrative when the combination of a mass-spring system with feedback control is introduced in the next section.

Another important aspect of the state-space model is the fact that the poles of the transfer function are equal to the *eigenvalues* of the  $\mathbf{A}$  matrix. These eigenvalues are those values of  $s$  where the determinant  $\det(s\mathbf{I} - \mathbf{A})$  equals zero, with  $\mathbf{I}$  being the unity matrix with the same dimension as  $\mathbf{A}$ .

$$\begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{c}{m} \end{bmatrix} = \begin{bmatrix} s & -1 \\ \frac{k}{m} & \frac{c}{m} + s \end{bmatrix} = 0 \quad (4.48)$$

When written out, this gives the following *characteristic polynomial* in  $s$ :

$$s\left(\frac{c}{m} + s\right) + \frac{k}{m} = s^2 + \frac{cs}{m} + \frac{k}{m} = 0 \quad (4.49)$$

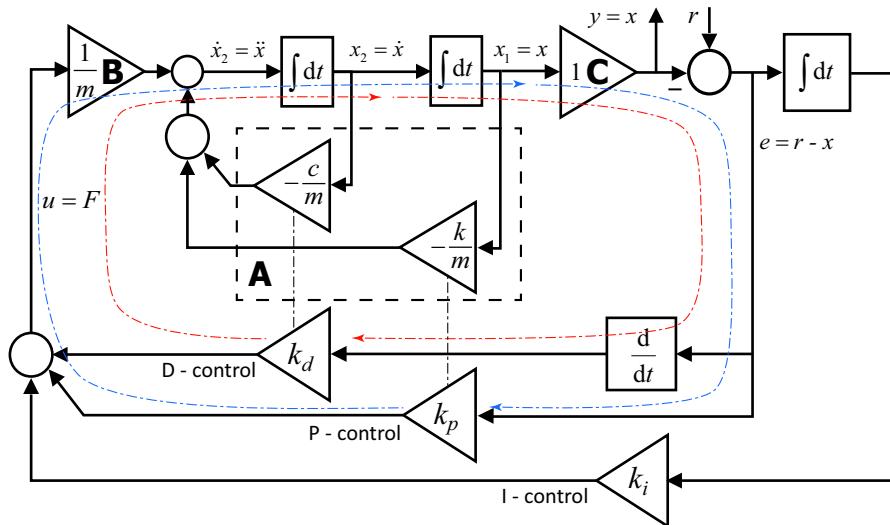
The poles are equal to the eigenvalue solutions of  $s$  of this polynomial:

$$p_{1,2} = -\sigma \pm j\omega_d = -\frac{c}{2m} \pm j\sqrt{\frac{k}{m} - \frac{c^2}{4m^2}} \quad (4.50)$$

With  $\omega_0 = \sqrt{k/m}$  these values of the poles become equal to the values that were found in Section 3.2.3.2 where the critical damping of a mass-spring system was derived:

$$p_{1,2} = -\sigma \pm j\omega_d = -\frac{c\sqrt{\frac{k}{m}}}{2\sqrt{km}} \pm j\sqrt{\frac{k}{m}}\sqrt{1 - \frac{c^2}{4km}} = -\frac{c\omega_0}{2\sqrt{km}} \pm j\omega_0\sqrt{1 - \frac{c^2}{4km}} \quad (4.51)$$

This means that by changing the terms in the  $\mathbf{A}$  matrix, the eigenvalues can be adapted such that the poles become located at a different position in the Laplace plane. This is exactly what happens with feedback control and the state-space model offers a very straightforward method to achieve an optimal pole location, as will be shown in the following examples.



**Figure 4.37:** Graphical state-space representation of SISO PID-control, applied to a damped mass-spring system. A feedback path, consisting of three terms, calculates the output of the controller  $u$  from the error  $e$  after differentiating and integrating. This representation shows that the blue dash-dotted loop over  $k_p$  has the same function as the internal loop by the stiffness  $k$  and the red dash-dotted loop via  $k_d$  has the same function as the damping coefficient  $c$  as the series of an integral and a differential term cancel each other out in the loop. The I-control action has no corresponding equivalent inside the uncontrolled system and is introduced to reduce the steady-state error. Note that the reference is inserted between the output  $y$  and the controller.

#### 4.4.1.2 PID-control feedback

To demonstrate the state-space approach with active feedback, the implementation in a PID-control setting is presented for a second-order mass-spring system.

A graphical state-space representation of the second-order mass-spring system under PID-feedback control is shown in Figure 4.37.

With this representation it is possible to trace the different feedback loops that work on the system separately. The two internal loops of the uncontrolled mass-spring system determine its dynamics with the spring and damper action. The three feedback control loops are added to these two internal feedback loops.

The blue dash-dotted P-control feedback loop over  $k_p$  runs parallel with

the internal stiffness loop with element  $-k/m$ . In combination with the input element  $1/m$  by **B**, the value of  $k_p$  is added to  $k$ . The P-control gain therefore corresponds to a modification of the stiffness of the system, just as presented in the example of the CD player. The similar relation exists for the differential gain  $k_d$  that corresponds to the damping coefficient  $c$  in the mechanical system. The red dash-dotted loop over the differentiator block ( $d/dt$ ) converts the measured position into a velocity that serves as the input to  $k_d$ . In the D-control feedback loop this calculated velocity is equal to the real velocity  $\dot{x}$ . In summary,  $k_p$  and  $k_d$  are used to change the stiffness and damping of the system.

A special word of attention has to be given to the reference input and the I-control loop with the additional integrator. In this PID-control example, the reference is indicated at the same position as where it was located with the analytical approach of the previous section on PID-control. With state-space control this location is more free to choose. In theory the reference signal can be inserted at any place in the control part because it is an externally definable signal. Often the reference is inserted at the output of the feedback controller. This is the same place where feedforward control is inserted, as was indicated at the beginning of this chapter in Figure 4.1.

The integration of the error introduces a third feedback loop. There is no equivalent element to the integrator in the uncontrolled second-order mass-spring system. In closed-loop, the action of the integrator results in a reduction of the steady-state error to zero and due to this effect it is sometimes called a *super-spring*.

Regarding the placement of the poles it is clear that the feedback constants directly determine the pole locations as  $k_p + k$  determine the undamped natural frequency with two poles on the imaginary axis while  $k_d + c$  shifts these poles over a circle back to the real axis. The fact that  $k_p$  and  $k_d$  are added to the terms in the **A** matrix indicates that another **A** matrix is obtained with different eigenvalues and this effect is further used with full state feedback.

#### 4.4.2 State feedback

Even though PID controllers still determine the mainstream of practical position control systems, it is useful to shortly introduce the more fundamental mathematical methodology, that enables to design control algorithms for complex systems with a higher number of inputs and outputs as well as system states. It is also a useful method when various positioning systems suffer from cross-couplings. This fundamental approach starts with the assumption that the best controller can be designed when the full model, including the dynamics and all states is known. By applying proportional feedback on the individual states, *state feedback* is realised. First this principle will be better explained and in the following sections it will be shown, how by means of system identification the state-space matrices can be determined and how state estimators are used to approach the ideal situation of all states being known as if they all are measured. Even in the case that direct measurement is not possible it will be shown that the states can be reconstructed (estimated) from a limited set of measured states.

The standard state-space notation of a dynamic system was shown to be equal to:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

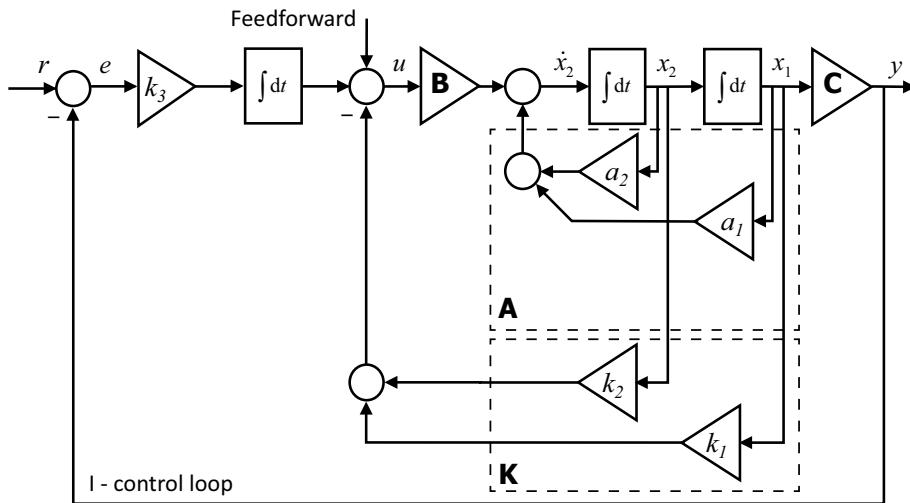
$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du}$$

In most systems the direct feed-through  $\mathbf{D}$  is not present or very small and as simplification  $\mathbf{D}$  is further assumed to be equal to zero for this explanation of state feedback. As explained, in state feedback it is assumed to have full information of the state vector  $\mathbf{x}$ . This means that all states of the system are directly measured or calculated from a limited set of measurements. The feedback-loop is closed with a feedback matrix  $\mathbf{K}$ , with  $n$  columns and  $i$  rows. The resulting output of the feedback controller equals  $\mathbf{u} = -\mathbf{Kx}$ . When the reference is not taken into account, as that only adds an additional setpoint term to  $\mathbf{u}$  without changing the closed-loop system dynamics, the state-space notation of the resulting dynamic system becomes:

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{B}(-\mathbf{Kx}) = (\mathbf{A} - \mathbf{BK})\mathbf{x} \quad (4.52)$$

$$\mathbf{y} = \mathbf{Cx}.$$

As a result a new system is created, looking the same as the open-loop system, but with a new eigendynamics matrix  $\mathbf{A}' = \mathbf{A} - \mathbf{BK}$ . So to achieve the desired performance of the closed-loop controlled system under state feedback, the desired pole locations have to be defined for stable operation and the required



**Figure 4.38:** State feedback on a SISO positioning system. The position  $x_1$  and velocity  $x_2$  are both measured. The derivative of the velocity,  $\dot{x}_2$  equals the acceleration and is in this case not measured nor required for full state feedback as this second order system is fully described by two states. By applying proportional feedback for the position and the velocity, the poles can be placed at the appropriate locations in the Laplace plane. The I-control loop is additional to reduce steady-state errors.

feedback matrix  $\mathbf{K}$  needs to be calculated. A state-space system is stable if the eigenvalues of the  $\mathbf{A}$  matrix have all a negative real part, directly corresponding to the poles of the system's transfer function. This means in a closed-loop controlled system, that all eigenvalues of  $\mathbf{A}' = \mathbf{A} - \mathbf{B}\mathbf{K}$  need to have a negative real part and sufficient damping for the imaginary (oscillatory) part in order to achieve a good control performance.

Figure 4.38 shows a graphic representation of a state feedback controller on a one dimensional (SISO) mass-spring positioning system, where the two states, the position and velocity, are measured and where the integral of the position is calculated. It is shown that the individual parameters  $k_{\#}$  of the feedback vector correspond to the individual parameters  $a_{\#}$  of the system's eigendynamics matrix. This is very similar to the case presented before for the PID controller of the second-order mass-spring system. This means that state feedback is equal to a set of P-control terms for a system where all states are directly measured, and the individual P-control gains correspond to the parameters of the feedback matrix  $\mathbf{K}$ .

In state-space control the integral action is a special case as one might say that this creates another state by the controller itself without any direct link to a similar state in the system. This state is not needed for controlling the system dynamics as two states are in principle sufficient for a second order system. As was previously explained, the I-control action is added to a feedback control system in order to reduce the steady-state errors by means of an increase of the loop-gain at low frequencies. For this reason it is customary in state-space control to introduce the I-control action as a separate negative feedback loop from the system output  $y$  to the input after comparing with the reference  $r$ . The integration will only reduce errors in the low-frequency area so high-frequency components of the reference signal will not be controlled and even attenuated by the I-control action ( $1/\omega$ ). This effect is solved by adding the high-frequency part of the reference to the feedforward input of the system. In that case the state-feedback loops will control this part of the frequency spectrum.

**Side note:** A critical designer of this control system also could say here: "Why should I use two sensors. One position sensor is enough as I can apply D-control to derive the velocity". While this statement is fully correct, one must not forget that it is always necessary to deal with sensor noise. If the position sensor is noisy, the derivative action would amplify this noise especially at high frequencies and this noise would be inserted in the feedback loop. For higher order systems with for instance multiple springs and bodies, where only the position of the last body is measured, even multiple derivatives of the sensor signal would be needed to apply full state feedback and this would amplify the noise even more. This is another example of the often economic trade-off that a mechatronic designer needs to make.

#### 4.4.2.1 System Identification

For a successful controller design good knowledge of the dynamics of the plant is crucial. Particularly for the design of model-based controllers a mathematical model of the uncontrolled system dynamics is required. But also for the tuning of classical controllers, such as a PID-controller, knowledge of the transfer function of the plant is necessary. These models can be obtained via physical modelling based on first principles, as described in several chapters in this book for the mechanical structure, the amplifiers and the actuators. An alternative to this physical modelling is to obtain a mathematical model by fitting a mathematical transfer function to the measured frequency response of the dynamic system via the *System Identification*

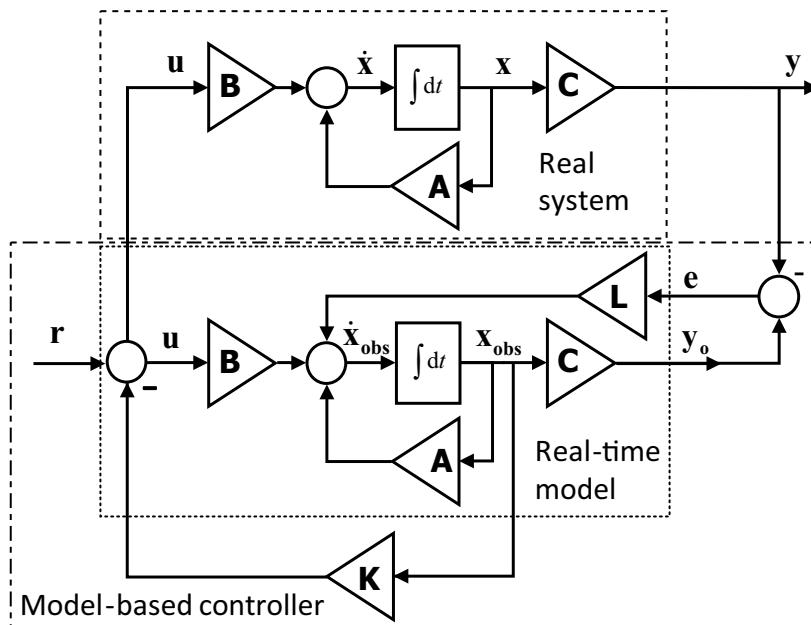
methodology.

With System Identification the model of the plant, including the sensor, is obtained by fitting the parameters of the differential equation to a measured set of input and output data. When the structure of the expected model is known, only the model parameters are identified, which is called *grey-box system identification*. If also the structure of the system is unknown, this is called *black-box system identification* and the identification algorithm then also has to estimate the model-order and structure. The system identification directly gives the state-space form or the transfer function of the system, but is always limited to the *minimal realisation* of the system, which means that from the system dynamics of the plant only those eigenmodes are considered that can be observed as well as controlled. This is in general not a problem when no other forces act on the system that excite the un-observed and un-controlled modes and it can even be desired as it avoids excitations of these modes by noise in the control loop. Nevertheless it is important to be aware of this fact because the un-observable or un-controllable eigenmodes might very well be a main cause for problems in a complex mechatronic system that consists of many actuated subsystems.

#### 4.4.2.2 State estimation

In the previous section it was shown how a suitable controller can be realised, if all states are measured. It is however not always possible to measure all positions, velocities and related derivatives in a mechatronic system, while precision sensors are also quite expensive. As mentioned before, using the derivative of the position signal in order to obtain the velocity signal, like in a D-control feedback system is possible, but may be too sensitive to noise. One way to get the full information of the state vector is to build a real-time estimator that is based on the model of the plant including the eigendynamics. Such an estimator is also called an *observer* while it observes the behaviour of a system by comparing it with the modelled behaviour. Such an observer also allows a trade-off between the bandwidth (speed) of the estimation and the noise performance. An observer with an optimal trade-off between these two important properties is called a *Kalman-filter*, named after the Hungarian mathematician and electronic engineer Rudolph Emil Kálmán.

Figure 4.39 shows the configuration of a state observer in combination with state-feedback control of the observed system. The blocks in the dashed box represent the real mechatronic system. The blocks in the dotted box represent the mathematical model that is implemented on a computer to



**Figure 4.39:** Model-based controller with an observer to estimate not measured states in a state feedback control system. The real-time feedback path is determined by the feedback matrix  $\mathbf{K}$  based on estimated state values from within the model. The model is updated by the difference between the observer output  $\mathbf{y}_o$  and the real system output  $\mathbf{y}$  via the matrix  $\mathbf{L}$ .

simulate the behaviour of the mechatronic system in real-time. When both systems receive the same input signal  $u$ , and both systems are identical, which means that a perfect model is available, both outputs  $y$  and  $y_o$  should be the same. However, in reality always modelling errors will occur while also the mechatronic system can be disturbed by external forces that are not taken into account and causing position and velocity errors. To compensate for these deviations the observer-gain matrix  $\mathbf{L}$  is introduced, that determines a feedback of the prediction error to the observer and is given by the difference between the output of the model and the output of the real system ( $\mathbf{e} = \mathbf{y}_o - \mathbf{y}$ ).  $\mathbf{L}$  has to be designed such that the closed-loop system for the observer part is stable. This is the case if the feedback loop of the observer, determined by  $\mathbf{A}$ ,  $\mathbf{L}$  and  $\mathbf{C}$ , only has poles in the left-half of the Laplace plane. The choice of  $\mathbf{L}$  influences the location of these poles, and determines the speed at which the state observer follows changes in the actual states of the physical system. When a fast response is chosen, the observed state

vector will quickly follow changes in the system but will also respond more to noise, originating for instance from the sensor. A slow response of the state observer introduces some “smoothing” or masking of fast disturbing signals, for instance when the measurement is corrupted by sensor noise.

Like with the name of an optimal observer, the observer gain that is providing the optimal trade-off between speed and noise performance for a known co-variance of the sensor noise is called the *Kalman-gain*.

#### 4.4.2.3 Additional remarks on state-space control

The analytical modelling of dynamic systems with transfer functions gets most attention in this book because it is more intuitively connected with the physical behaviour of a mechatronic system. Nevertheless the state-space approach has several advantages over the analytical transfer function approach especially with more complex systems. The matrix calculations enable to directly create a MIMO controller where all cross-coupling between different channels is taken into account. This second advantage entails also a disadvantage as it becomes more difficult to distinguish problems that are caused by malfunctioning of one element in the complex system. This argument has long hampered the application of MIMO controllers in practical mechatronic systems.

Furthermore the state-space model allows to more easily include any initial conditions and additional disturbances that act on the same system states.

The observer based controller has another potential advantage. While the standard implementation of an observer runs on all measurements in real-time at the same sampling rate as the controller, it is also possible to decouple these sampling rates. The time behaviour of the  $\mathbf{L}$  matrix can be chosen in relation to the level of fidelity of the model to reality. When the model is almost ideal, the estimation error  $\mathbf{e}$  will show very little deviations and when these deviations are slow,  $\mathbf{L}$  can be reduced to correct the model only occasionally. This introduces the possibility to decouple the timing of the measurement path from the timing of the feedback loop.

One can think of several situations where measurements can only be done at certain long intervals. One example occurs when an object has to be placed at another location for the measurement because of environmental constraints. Another example is the situation where the process needs to be terminated for a precise measurement because the process induces an excess of noise and disturbances. In those cases, the observer provides information to the feedback loop prior to the real measured value, so running rather as a

simulation of the dynamics, thus preventing instability by the phase lag of the slow measurement signal. As long as the measurement delay is known and the model is sufficiently accurate, this method can be really effective as it defines a process of frequent re-calibration of the real-time dynamic model at regular intervals. One might also say that at regular intervals the real feedback loop is temporarily broken and between these intervals, the observer will act as a simulator or as an adaptive feedforward controller. When the model is not perfect, a combination of real time local approximate measurements with observer based correction at longer intervals on the overall performance might still give the required result. The system would behave like a ship sailing over the ocean in ancient times, where the captain as controller determined his direction by real-time feedback corrections that were based on the observation of the wind and the ocean waves, while he corrects his model by infrequently "shooting a star" with his sextant.

## 4.5 Limitations of linear feedback control

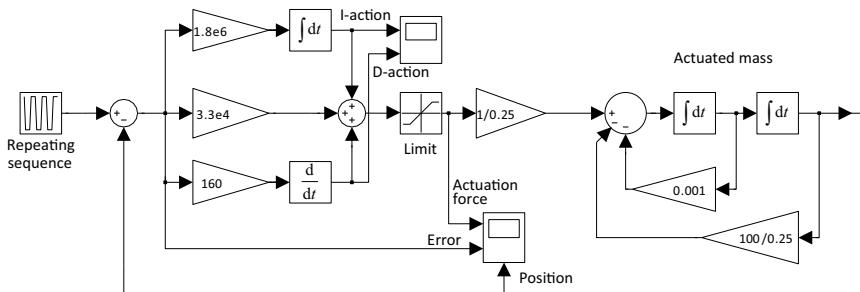
As stated in the introduction of this chapter, for simplicity only linear (control) systems are discussed in this book. One of the main limitations of linear control is the assumption that the system states of the plant are possible and can be measured over the total range of interest. Further it is assumed that the control action is not limited. In reality any linear system will have its boundaries such as range limitations and other non-linearities. For the sake of completeness this last section deals with some of the most common non-linearities that occur in precision positioning systems. The discussion on how to compensate or account for these possible limitations by the design of an appropriate (non-linear) control system would result in a book that is about as comprehensive as the one that you are reading now. For that reason non-linear control is not included here even though non-linear control carries a good potential to improve system stability and push the performance-limits even higher. For further information on this topic it is advised to read the control-related references that are listed in the appendix.

When taking a detailed look again at the block diagram of the motion control system shown in Figure 4.1 it is possible to identify some potential sources of non-linearities.

On the side of the plant any physical system certainly will have its limitations, such as a maximum measurement range of a sensor, called *clipping* and the limitation of the actuation range of actuators by *saturation*. For instance piezoelectric actuators that are based on a stiff design can only move a few micrometres. Other examples of limitations are the maximum force that a Lorentz actuator in a zero-stiffness design actuator can generate. In the amplifiers any *slew-rate* limitation of the speed of change of the output and maximum current or voltage of the power stage will have impact on the system performance. Further to be mentioned are non-linearities like backlash, friction, and non-linear stiffness in the mechanical components or creep and hysteresis as they occur for instance in the piezoelectric transducers that are presented in Section 5.6 in the next chapter.

Also in the (digital) control system sources of non-linearities and limitations can be identified, such as the quantisation (in value) and sampling (in time) of the continuous sensor signal by the A/D-converter, the limited update rate and resolution of the control signal due to the quantisation of the D/A converter, but also timing uncertainties in the sampling of the converters, the so-called jitter.

While the limitations due to the control system can easily be accounted for

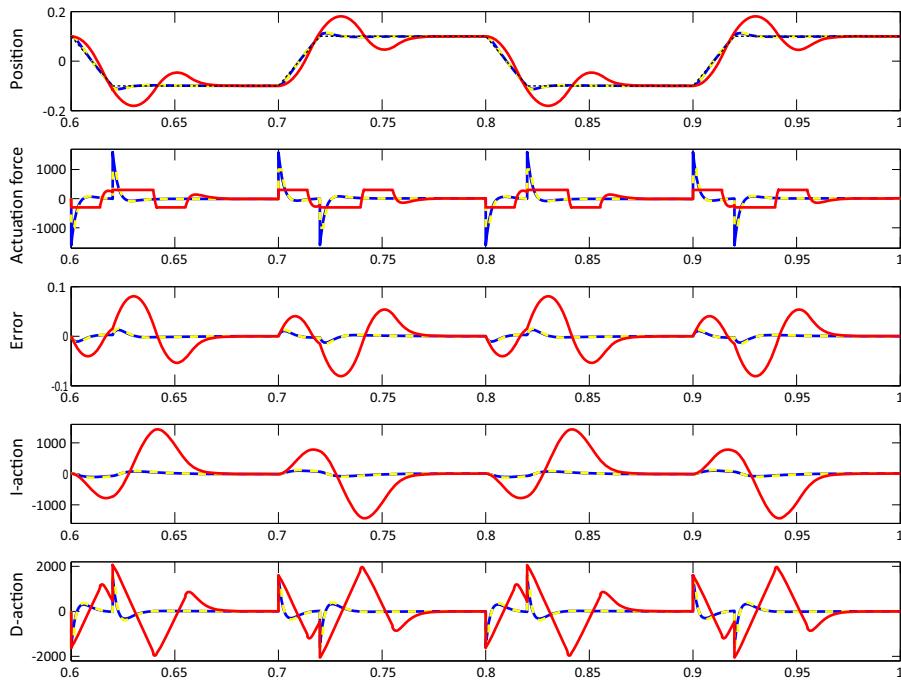


**Figure 4.40:** MATLAB-Simulink model of the PID controlled mass-spring system of Equation. 4.33 with saturation of the actuation force.

by choosing the right components, limitations that are present in the plant are a frequent source of errors and can even cause damage to a system when being overlooked. One reason for such errors in practice is that standard modelling software packages, such as Matlab and Simulink, are allowing extremely large control action, error inputs, and system states, that may not have any relation with capabilities of the real system. The limitations (saturation) of actuators and sensors have to be taken into account and must be integrated in the Simulink-models by adding appropriate function blocks, like shown in Figure 4.40, in order to keep the model close to the actual physical system. This means, that a practical insight in the limitations of sensors, amplifiers and actuators is necessary to prevent costly mistakes in modelling mechatronic systems.

As an example to visualise the effect of such limitations, Figure 4.41 shows the behaviour of the PID-controlled system from the previous section (see Equation 4.33) in case of a saturation of the actuation force, putting a limit to the maximum possible control effort. The reference signal is a trapezoidal signal with a fundamental frequency of 5 Hz and a peak-to-peak amplitude of 0.2 m (dotted black line in the first panel of the figure). For the unconstrained system the actuation force ranges up to about 1600 N (blue solid line), while the control effort for the constrained system is limited in the simulation to 1 kN (yellow dashed line) and 300 N (red solid line), respectively.

While the control action exceeds the actuation force limit, the control-loop is broken and the system does not react to changes in the control action as the maximum actuation force is already applied (and with this the maximum energy that can be supplied to the system per unit of time). This means that due to this saturation the linear control system becomes non-linear



**Figure 4.41:** Behaviour of a mass positioning system under PID-control with varying limitation of the maximum actuation force: unconstrained (blue line), limited to 1000 N (dashed yellow line), and limited to 300 N (solid red line). The first plot shows the controlled position of the mass in response to a trapezoidal reference signal (black dotted line). The second plot shows the output of the PID-controller, equalling the actuation force. The third plot denotes the error signal, which with a scaling factor also equals the P-action. The fourth and fifth panel show the I-action and D-action. The mismatch between the actuation expected by the linear (unconstrained) PID-controller, given by the blue lines, and the actual actuation causes a significant degradation in control performance for the case where the actuation force is limited to 300 N, and the system starts to fiercely oscillate.

and the system does no longer react as the PID-controller would expect. For the case where the actuation force is limited to 1 kN, the saturation level is reached only for a very short moment (dashed yellow line) due to the D-gain when the reference signal changes. This is one of the reasons why for practical systems one should use smooth reference signals rather than steps and impulse-like signals. As this saturation is limiting the control action only very little, the increase of the control error is also very small

and the system remains in saturation only slightly longer for the time that the unsaturated force would exceed this value of 1 kN. When the system re-enters the linear regime, the PID-controller easily can compensate for this slightly larger error during saturation.

For the case where the actuation force is limited to 300 N, however, the energy per time unit that can be put into the system for control is much lower than the PID-controller would demand for the given reference signal. The mismatch between the linear system and the actual system is too large and leads to a larger control error and a significant degradation of the system performance. The main problem here is, that the I-gain of the controller continues to accumulate the error during the entire time that the system is in saturation, which takes significantly longer and consequently results in much larger accumulation of the I-action as would be the case for the unconstrained system. This effect is called *integrator windup*. This longer integrated and larger error cannot easily be compensated when the system re-enters the linear regime, as the I-action only starts to decrease once the error changes its sign like when the system is moving in the opposite direction. This results in strong oscillations of the system and even may lead to *limit-cycling* as the too large accumulation of integral action can lead to a bang-bang action where the control signal is just switching between the positive and the negative saturation of the actuation force.

In real systems measures have to be taken in the controller to prevent these phenomena. One of the possibilities is to reduce or even cancel the I-control action until the system is in the linear range again within the saturation limit of the actuation force. This is called *anti-windup control* and should be present in all PID-control systems, and is often sufficient when long range sensors are used.

This example demonstrates that linear control, as powerful as it is, also has its limitations that have to be known and considered in the design of the control system.

Another limitation, not shown in the figure, would be the limited measurement range of a sensor. When the sensor is out-of-range, the control-loop is not broken towards the actuator but the controller is "flying blind" once the sensor signal clips. This means that the actual error may be much larger than the controller observes, but does not get compensated. With very short range sensors and long range actuators, like in a close tracking system, one can reduce the P-gain and increase the D-gain until stability is achieved. The auto focus of a photo camera is a good example of such a problem, where often the capture range of the focal sensor is much smaller than the focal

setting range of the lens and outside the range no information is available about the direction. As a result the camera keeps scanning past the sharp spot without stopping in time.

As a last example for the limitation of linear control backlash needs to get some attention as this non-linearity is often present in less advanced mechatronic system where the cost of direct drive actuators and air bearings is prohibitive. With a positioning system that consists of a rotary DC motor with a gearwheel transmission, the backlash and friction of the transmission disturbs the transfer between the motor and the positioned mass. This non-linearity limits the possibility to achieve a fast and well damped feedback controlled system with precise positioning of the mass because during reversal the transmission is broken by the backlash. This problem can be partly solved by mounting a *tacho-generator* directly at the DC-motor shaft. This sensor provides reliable state information about the velocity of the motor and can be used for stabilisation of the (high-bandwidth) inner position/rotation feedback loop. Non-linear errors of the transmission will however not be measured by the tacho-generator and as a consequence they will not be reduced in the positioning of the load-mass.

This section of the book should for sure not be misunderstood as linear control is extremely powerful and important in motion control systems. In most applications the performance of linear control is however limited by the boundaries of the physical plant and these have to be considered in the controller design. As stated above we would like to refer to the references on linear and non-linear control given in the appendix for further reading, in order to get an insight how modern non-linear control methods may allow to achieve a control performance beyond the limits of linear control.

## 4.6 Conclusions on motion control

In this chapter it was shown that under suitable circumstances, feedback and feedforward control enable to realise a significant improvement in the dynamic performance of mechatronic motion systems. Feedback control allows to modify the system properties by changing the pole locations of the system, therefore offering to control unstable systems and add robustness to the feedback controlled system. Feedforward control enables to improve the performance of motion system for instance by zero-pole cancellation for reference following, while not being limited by the conditions for stability and, in general, being simpler and faster than feedback control.

With the combination of both, feedforward and feedback control, also called *two degree of freedom control*, the system designer can optimise the control design including the trade-off between performance and robustness.

It was also demonstrated that active control requires reliable information of at least one of the system states. Also the behaviour of the plant has to be known sufficiently well, either by means of system identification or by modelling. In the following chapters the behaviour of the actuation and sensing elements are presented to enable modelling of these important elements of the mechatronic plant. The final chapter about wafer scanners will further underline the learned theory by illustrating the importance of both feedforward and feedback control with practical data.

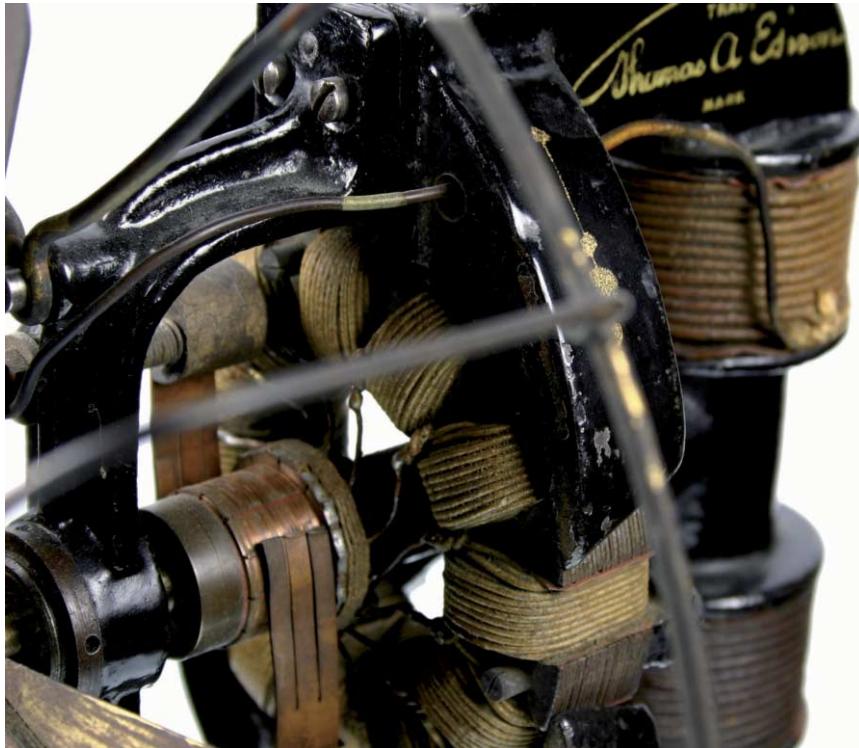
# Chapter 5

## Electromechanic actuators

The eigen dynamics of a mechatronic system are not only determined by its mechanical dynamic performance. Like mentioned before, the incorporated subsystems such as amplifiers, actuators and sensors have their own impact on the full system dynamic performance. It requires a thorough understanding of the interaction between the mechanical and electrical engineering domains regarding energy conversion, in order to design a well balanced system. This interaction is covered in this chapter on electromechanics. Although the subject of electromechanics describes the transition from electrical to mechanical engineering domains and vice versa, this chapter will only cover the part that is related to actuation. The reversed phenomena will be presented in Chapter 8 on measurement.

After a general introduction, the physics that are related to electromagnetic forces is covered including the applied Maxwell equations and the calculations on magnetic fields. Starting with Section 5.2 the behaviour of linear actuators like moving-coil Lorentz actuators and the stronger but non-linear reluctance actuators will be presented together with the optimal combination of both in the hybrid actuator. These basic configurations are sufficiently representative for the design of most electromagnetic actuators. With the derived simplified equations, reasonable calculations can be made on concept designs. Further attention is given to the electrical properties, that are important for the integration in mechatronic systems, especially related to the amplifier that has to deliver the current. Electromagnetic actuators still determine the main-stream actuation principle in precision

mechatronic systems, but piezoelectric actuators are increasingly important especially in applications that need very large forces with small corrective movements, like in machining tools. But also in very small systems piezoelectric actuators are used to advantage, because they can create displacements at very high frequencies. For this reason, Section 5.6 presents an introduction on piezoelectric actuators.



**Figure 5.1:** Electric motor that was designed by Thomas Alva Edison to drive a cooling fan. The motor contained no permanent magnets while the magnetic flux was only created by current. A commutator with sliding contacts allowed it to be used with direct current electric power that was his preferred source of electrical energy.

(Courtesy of John Jenkins, [www.sparkmuseum.com](http://www.sparkmuseum.com))

## 5.1 Electromagnetics

An *electromagnetic actuator* is a special type of *electric motor*. The term actuator is reserved for those devices that directly convert electrical energy into a movement over just a limited linear or rotational range. Before presenting the theoretical background on electro magnetics first a bit of history is presented.

### 5.1.0.4 History on magnetism

The first experiments regarding electricity by the American statesman and scientist Benjamin Franklin (1706 – 1790) occurred already around the year 1750. In spite of these early investigations the first real application, the creation of artificial light, only emerged in the second half of the 19th century. Also in this period the first electric motors were created based on the experiments on electromagnetic forces by the Danish physicist Hans Christian Oersted (1777 – 1851). An example of such an early electric motor is the one shown in Figure 5.1 designed by Edison around 1898 for use in a cooling fan. Thomas Alva Edison (1847 – 1931) was an American inventor and scientist but above all he was a successful businessman who did everything he could to increase interest for the use of electricity. Next to his well-known contributions to the application of incandescent light, he also made other electric appliances like the electric fan of which the electric motor is shown here.

For a long time the rotating electric motor determined all electrical driven motion in the world. These motors worked according to a variety of principles ranging from asynchronous induction to synchronous permanent magnet types, with or without commutation, that enabled them to run at speeds in the order of several thousands revolutions per minute. The application of gearboxes made it possible to convert their usually high rotation speed into a slower and more controlled motion. Next to a rotating motion also linear motion was possible with crown wheel and screw spindle mechanisms. In the course of the 20th century already high levels of linear precision in the order of ten to hundred micrometres could be achieved, with the use of pre stressed (ball-screw) mechanisms. These were mainly used in machining tools, but people started to look for alternatives that did not suffer from the inherent backlash, friction and play of these mechanical drives, that limited the use with the “servo” active position control systems that were just introduced at that time. As discussed in the introduction of this book this

has resulted in two different directions. The first originated in the machine tool industry and mostly used the newly discovered piezoelectric properties of certain materials to use them as actuators. The second direction was based on positioning of systems without strong external forces and relied mainly on electromagnetic energy conversion principles.

### 5.1.1 Maxwell equations

Electro magnetics is the physics area that describes the phenomena, associated with electric and magnetic fields and their interaction. Although this book focuses on comprehension, rather than mathematics, it is not possible to escape from the fact, that the physics in electro magnetics are governed by stipulated laws and models. These laws in itself can not be understood by definition, but have to be accepted for reason of their capability to predict the behaviour of electromagnetic systems. Though this complicates the matter for people, who want to know why things work as they do, these laws are taken as a starting point and then the subject is presented as if its understood what causes these laws to be true.

These basic laws are called the “Maxwell equations” and will be presented in the following to be used in the further practical implementation.

The Scottish physicist and mathematician James Clerk Maxwell (1831-1879) formulated his equations partially based on the work of previous scientists, but the laws got his name because of the way he combined them. The Maxwell equations are a set of four equations, with the status of physical laws, stating the relationships between the electric and magnetic fields and their sources being charge density and current density.

Table 5.1 defines the variables that play a role in this chapter. For a complete overview also the electrical variables, that were defined in Chapter 2 are mentioned. Like the previously defined electric field **E**, the electrical current density **J** is a vector field representing the movement of free, or *unbound* charges in space. The electric permittivity  $\epsilon_0$  has also been previously defined and the electric charge density  $\rho_q$  represents the amount of unbound charges in a volume.

The magnetic field **B** is a vector field. Like with the electric field, it is graphically represented by the density and direction of magnetic flux lines. In the design of electromagnetic actuators, the magnitude of the magnetic field in the direction of interest is called the flux density  $B$ . It directly relates to the quantitative nature of the magnetic flux  $\Phi$  in that direction, calculated by integrating the magnetic field over a surface perpendicular to

**Table 5.1:** Physical variables in electromagnetism.

Physical quantity	Symbol	SI unit
Electric field	<b>E</b>	[V/m]
Electric current density	<b>J</b>	[A/m <sup>2</sup> ]
Electric charge	<i>q</i>	[C]
Electric permittivity in vacuum	<i>ε<sub>0</sub></i>	[As/Vm]
Electric charge density	<i>ρ<sub>q</sub></i>	[C/m <sup>3</sup> ]
Magnetic field	<b>B</b>	[T]
Magnetic flux density	<i>B</i>	[T]
Magnetic flux	<i>Φ</i>	[Wb]
Magnetic permeability	<i>μ</i>	[Vs/Am]
Magnetising field	<b>H</b>	[A/m]
Magnetic field strength	<i>H</i>	[A/m]

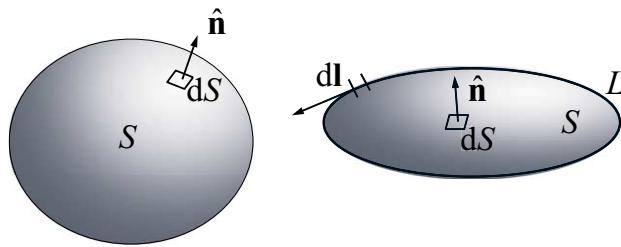
that direction.

The magnetic permeability  $μ$  represents the possibility of a magnetic field to pass through an object. The word permeability stems from the verb “to permeate”. The permeability in vacuum is called  $μ_0$  and is a reference for the permeability of other materials.

The magnetising field **H** and its related magnitude, the magnetic field strength *H*, are directly connected to the magnetic field by the magnetic permeability:  $\mathbf{B} = μ\mathbf{H}$  and  $B = μH$ .

The Maxwell equations can be written either in the integral or differential form. The integral form is more easy for explaining the meaning and will be used to derive the equations in this chapter. The differential form is better suited for mathematical modelling. In order to keep this differential form more easy to write down, the mathematical terms *divergence* (div) and *rotation* (rot) are used. These are respectively the dot (div) and cross (rot) product between the differential vector nabla ( $∇$ ) and the vector of interest. In the three dimensional space  $∇$  equals:

$$\nabla = \left[ \frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right] \quad (5.1)$$



**Figure 5.2:** Definitions of the vectors  $\hat{\mathbf{n}}$  and  $d\mathbf{l}$ , surface  $S$  and closed loop  $L$  that are used for the Maxwell equations. The left drawing shows surface  $S$ , that encloses a volume and is used with both Gauss's laws. The right drawing shows the closed loop  $L$  with its enclosed surface  $S$  and is used with Faraday's and Ampère's law.

With the previously defined variables and using the definitions of Figure 5.2, the first Maxwell equation is written as follows:

- *Gauss's law (electric):*

$$\iint_S (\mathbf{E} \cdot \hat{\mathbf{n}}) dS = \frac{q_{\text{enc}}}{\epsilon_0} \quad (5.2)$$

$$\text{div } \mathbf{E} = \nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (5.3)$$

This first law was originally postulated by the German mathematician and scientist Johann Carl Friedrich Gauss. It states that the surface-integral of the electrical field over any closed three dimensional surface  $S$ , like for instance a sphere, equals the charge  $q_{\text{enc}}$  enclosed within the closed surface, divided by the electric permittivity ( $\epsilon_0$ ). It is directly related to Equation (2.2) of Chapter 2. With electromagnetism this law is not used, because in electromagnetic actuators all electrical charges are bound, which means that they are not free and always in equilibrium with the positive charge of protons of the wires that are used to carry the current. Nevertheless it is mentioned for completeness and also to demonstrate the difference between electric fields and magnetic fields as given in the second Maxwell equation.

- *Gauss's law (magnetic):*

$$\iint_S (\mathbf{B} \cdot \hat{\mathbf{n}}) dS = 0 \quad (5.4)$$

$$\text{div } \mathbf{B} = \nabla \cdot \mathbf{B} = 0 \quad (5.5)$$

This second law of Gauss states that the surface-integral of the magnetic field over a closed surface  $S$  is always zero. With any closed surface, the magnetic flux entering the volume within the closed surface is equal to the magnetic flux that exits that volume. When represented by flux lines this can only be true when all flux lines form a closed loop. Gauss's law on magnetic fields is based on the observation that magnets always act as dipoles, a north and south pole where the flux flows internally from south-to north pole and externally back from north- to south pole. This is the main difference with Gauss's law on electric fields where electrical charges can exist without a monopole counterpart. In case of a single charged particle, the equivalent electrical field lines originate in the charge and just go to infinity. Integrating over a surface surrounding that charge would result in a finite value according to the first law of Maxwell. Insight and understanding of Gauss's law on magnetic fields is necessary for calculating the flux in magnetic circuits.

The third Maxwell equation gives the relation between a change in the magnetic field and the resulting induced electrical potential difference in a wire that surrounds that field.

- *Faraday's law:*

$$\oint_L \mathbf{E} \cdot d\mathbf{l} = - \frac{d}{dt} \iint_S (\mathbf{B} \cdot \hat{\mathbf{n}}) dS \quad (5.6)$$

$$\text{rot } \mathbf{E} = \nabla \times \mathbf{E} = - \frac{\partial}{\partial t} \mathbf{B} \quad (5.7)$$

This law was originally postulated by the English chemist and physicist Michael Faraday (1791 – 1867). It states that the line-integral of the electrical field over a closed loop  $L$  equals the change of the flux through the open surface  $S$  bounded by the loop  $L$ . In Chapter 2 the electromotive force  $\mathcal{F}_e$  was introduced as this integral over the electric field inside a voltage source. For this reason the third Maxwell equation is the foundation under the theory that describes the creation of electricity by magnetism. A voltage source is created inside the windings of a coil by a changing magnetic field surrounded by that coil. It explains also the phenomenon of self-inductance, as will be defined in Section 5.3. An important element in this law is the minus sign that indicates, that the direction of the electric field is opposite to the vector  $d\mathbf{l}$ . It explains several properties of electromagnetic actuators that will be presented in this chapter, including the damping effect that is observed when the actuator is supplied from a low impedance source.

The conversion from mechanical into electric energy implicitly also leads

to the reverse effect, which means that by inducing a current through an electric wire within a magnetic field this wire will experience a force, the *Lorentz force*, that will be presented further on in this chapter.

The fourth Maxwell equation gives the principle of the creation of a magnetic field by an electric current.

- *Ampère's law:*

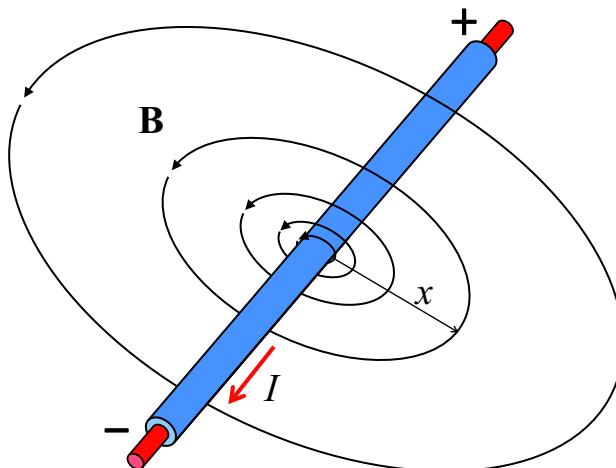
$$\oint_L \mathbf{B} \cdot d\mathbf{l} = \mu_0 I + \epsilon_0 \mu_0 \frac{d}{dt} \iint_S (\mathbf{E} \cdot \hat{\mathbf{n}}) dS \quad (5.8)$$

$$\text{rot } \mathbf{B} = \nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \epsilon_0 \mu_0 \frac{\partial}{\partial t} \mathbf{E} \quad (5.9)$$

This law, originally postulated by Ampère, states that the line-integral of the magnetic field over a closed loop  $L$  is equal to the sum of two terms. The first term represents the current that flows through the opening of the loop and the second term represents the change of the electric field over the surface that is enclosed by the loop. This second term is in reality not relevant for electromagnetic actuators, again due to the bound character of the charges as mentioned with Gauss's law on electric fields. This means that in a reduced form, without the unbound charges, this law gives the relation between the magnetic field and the current through a wire. It is used in modelling magnetic fields induced by a current in a coil.

### 5.1.2 Magnetism caused by electric current

The Maxwell equations in their differential vectorial notation are applied in many finite element modelling software packages. The resulting calculations reach near perfection, depending on the refinement of the chosen grid, the exactness of the shape of the magnetic parts and the affordable amount of computing time. For a mechatronic system designer however, it is often necessary to quickly obtain a first estimation of the properties of an electromagnetic system, before entering into detailed analysis with dedicated software. It is often also necessary to be able to recognise artifacts that are created by the modelling software when the boundary conditions are badly chosen. For that reason in the following sections the vectorial notations are often approximated by scalar equations dealing with magnitude only, assuming the direction of the fields and currents is known. This assumption is only valid when the directional correlation between all parameters are either parallel or orthogonal. This first order approximation is often allowed



**Figure 5.3:** The magnetic field, generated by a current through a wire is directed clockwise, when looking in the direction of the current. The magnitude of the magnetic field decreases proportional with the distance to the wire.

in initial actuator designs, because of their orthogonal design. When necessary an angular factor (sine/cosine) can be introduced as a second order correction.

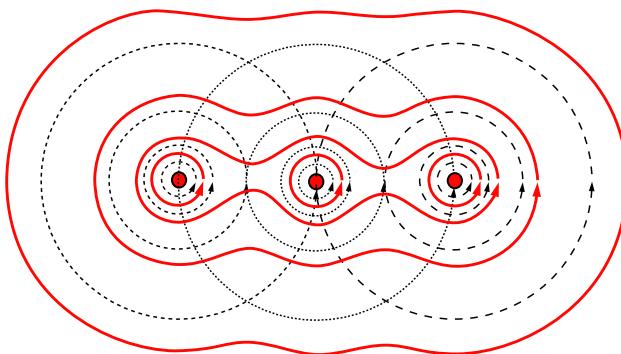
The steps toward usable equations start with the basic relation between electricity and magnetism derived from Ampère's law without the term for the unbound charges:

$$\oint_L \mathbf{B} \cdot d\mathbf{l} = \mu_0 I \quad (5.10)$$

When the loop is a circle with radius  $x$  around a wire with a current  $I$  as shown in Figure 5.3, the magnetic field  $\mathbf{B}$  has a constant flux density  $B$  over the circle and directed along the circle. This means that the left integral term becomes equal to  $2\pi x B$ . As a result the flux density becomes equal to:

$$B(x) = \frac{\mu_0 I}{2\pi x} \quad (5.11)$$

This equation is valid under the condition that the surrounding material is vacuum, for which Ampère's law was postulated. In words this equation tells us, that the magnetic flux density is proportional to the current in the wire and the magnetic permeability in vacuum and inversely proportional to the distance. This dependence of the magnetic permeability automatically



**Figure 5.4:** Three currents running in parallel wires, directed towards the observer, each generate a magnetic field that is represented by the dashed field lines. These fields add vectorial, according to the superposition principle of magnetic fields, resulting in a larger total field that is represented by the solid red field lines. It shows the straightening of the resulting field lines at some distance parallel to the orientation of the wires.

results in a change of the flux density when the permeability would be different.

More directly related to the electric current is the magnetising field  $\mathbf{H}$  with its scalar magnitude, the magnetic field strength  $H$ :

$$H(x) = \frac{B(x)}{\mu_0} = \frac{I}{2\pi x} \quad (5.12)$$

This means that the magnetic field strength is only determined by the current and the distance to the wire. For this reason the vector field  $\mathbf{H}$ , corresponding with the magnetic field strength  $H$  was named “magnetising field” in order to indicate an electric current as the source of the magnetic field  $\mathbf{B}$ <sup>1</sup>.

In practical electromagnetic systems wires are wound in coils with a certain number of windings  $n$ . As can be seen in Figure 5.4 the magnetic fields created by a multiple of electrical wires running in parallel, add vectorial, because of the superposition principle of magnetic fields. As a result, by winding the wire in a coil, the total field generated by the current increases per winding added. After integration of the magnetic field of all windings using Equation (5.11) over the cross section  $A_c$  inside a coil with a winding

---

<sup>1</sup>The real answer to the philosophical question of “what comes first” or “what causes what”, the  $\mathbf{H}$  or the  $\mathbf{B}$  field, is beyond the needs of this chapter as both fields are equally present and can be used where appropriate.

height  $h_c$ , the following expression for the flux density ( $B_w$ ) and magnetic field strength ( $H_w$ ) inside the coil of Figure 5.5 would be obtained in vacuum or air.

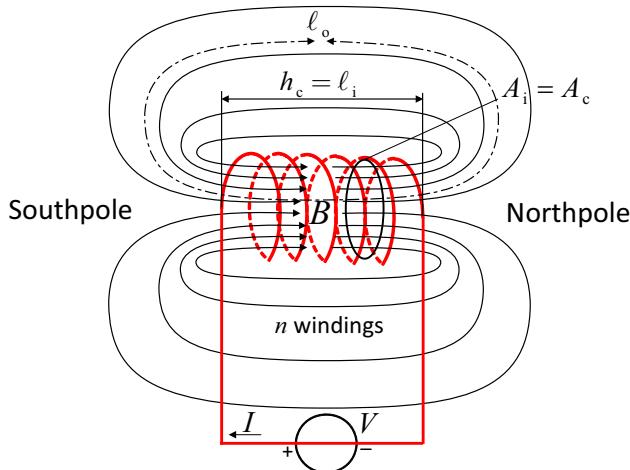
$$B_w \approx \frac{\mu_0 n I}{h_c} \quad \Rightarrow \quad H_w = \frac{B_{w,i}}{\mu_0} \approx \frac{n I}{h_c} \quad (5.13)$$

This relation is a rough approximation because of the fact that the magnetic field is not really homogeneous. The effects of the edges of the coil are fully neglected, which limits this approximation to coils with a height to diameter ratio of larger than one. Nevertheless it is useful in a qualitative manner, as it shows the impact of the height of the coil. The reduction of the magnetic field as function of the winding height  $h_c$ , under further equal conditions regarding current and number of windings, is caused by the reduced addition of the separate flux levels of each winding at an increased mutual distance. The diameter is cancelled out of the equation, again under the condition, that it is approximately the same value or smaller than the height. This can be reasoned from the fact, that in the mid position the contributions of the different windings is equal, while closer to one side the contribution of that side is proportionally increased, where the contribution of the other side is decreased. In approximation this means, that the flux density is constant over the cross section.

The direction of the magnetic flux, relative to the current direction, directly follows from the definitions for the Maxwell equations from Figure 5.2. Following the direction of the current, the magnetic flux is rotating clockwise around the wire as was shown in Figure 5.3. For a coil with multiple windings this means that the direction of the flux inside the coil corresponds with the current flowing around it in the clockwise direction. The place where the flux leaves the magnet is defined as the “North pole” and the other side as the “South pole”. Even though this seems to correspond with their counterparts of the magnetic field of the planet Earth, the poles of the earth are reversed as the definition of North pole in a magnet was made to tell which side of a compass needle points towards the North pole of the planet Earth. Because opposite poles attract each other this definition has caused the reversal.

### 5.1.3 Hopkinson’s law

With the found relations between the current in a wire and the generated magnetic field it is possible to derive a simplified model, that is useful for a first order analysis of magnetic systems. Similar to Ohm’s law in electronics,



**Figure 5.5:** The magnetic field from a current carrying coil consists of the vectorial addition of the magnetic fields of all windings contributions. This results in an approximately uniform field inside the coil that is shared by all windings.

the British physicist and electrical engineer John Hopkinson (1849 – 1898) stated that a magnetic system can be described as a combination of a source of magnetism that, together with a magnetic load, determines the magnetic “current”, the flux  $\Phi$ .

It was shown in the previous section that an electric current acts as a source of magnetism. In Hopkinson’s law of magnetics this source is defined as the multiplication of the current with the number of windings. It is called the *magnetomotive force*, familiar to the electromotive force for a voltage source of electricity:

$$\mathcal{F}_m = nI \quad [\text{A}] \quad (5.14)$$

For the magnetic resistance the term *Reluctance* ( $\mathfrak{R}$ ) is introduced because of the English word “reluctant” in relation to the unwillingness of a material to be permeated by a magnetic field. The magnetic reluctance is defined by means of the following relation:

$$\mathfrak{R} = \frac{\ell_\Phi}{\mu A} \quad (5.15)$$

where  $\ell_\Phi$  equals the length of the path, that the magnetic flux has to follow and  $A$  equals the cross section of the path. This is logical, when considering that the reluctance  $\mathfrak{R}$  represents the amount of effort that is needed to

introduce a magnetic flux in a certain material. A larger surface and higher values of the permeability will reduce this effort while a longer path length will increase it.

With these definitions, Hopkinson's law of magnetics, similar to its electric counterpart, is formulated as follows:

$$\Phi = \frac{\mathcal{F}_m}{\mathfrak{R}} = \frac{nI\mu A}{\ell_\Phi} \quad (5.16)$$

Generally neither the cross section  $A$ , the path  $\ell_\Phi$  nor  $\mu$  is constant over the entire system, so in practice the reluctance is calculated as a summation over the path of the flux of the reluctance contributions of each region with different properties. This means that each reluctance element adds to the total reluctance just like with their electrical counterparts.

To check this law the induced magnetic field by a current in the coil from Figure 5.5 in vacuum is determined, using Hopkinson's law of magnetics. This results in a flux density  $B_w$  inside the windings of the coil of:

$$B_w = \frac{\Phi_w}{A_c} = \frac{\mathcal{F}_m}{A_c \mathfrak{R}} = \frac{nI}{A_c \mathfrak{R}} \quad (5.17)$$

The reluctance can be modelled to consist of the part  $\mathfrak{R}_i$  with length  $\ell_i$  inside the coil and the part  $\mathfrak{R}_o$  with length  $\ell_o$  outside the coil:

$$\mathfrak{R} = \mathfrak{R}_i + \mathfrak{R}_o = \frac{\ell_i}{A_i \mu_0} + \frac{\ell_o}{A_o \mu_0} \quad (5.18)$$

where  $A_i$  and  $A_o$  equal the cross sections of the flux path inside and outside of the coil. The first term of the reluctance inside the coil can be determined quite straightforward, because of the constant cross section  $A_i = A_c$  and the height of the coil ( $\ell_i = h_c$ ). The second term of the reluctance outside the coil is more complicated, as the length  $\ell_o$  for each flux line will be different ranging from  $h_c$  at the exit of the coil to infinite. Fortunately the cross section  $A_o$  is also infinite, as it consists of all space outside the coil. As a result the reluctance of the outside part becomes small in respect to the inner part. This means, that the reluctance can be approximated by only taking the inner part:

$$\mathfrak{R} \approx \frac{\ell_i}{A_c \mu_0} \quad (5.19)$$

The flux density and related field strength become:

$$B_w \approx \frac{\mu_0 n I}{\ell_i} \approx \frac{\mu_0 n I}{h_c} \implies H_w \approx \frac{n I}{\ell_i} = \approx \frac{n I}{h_c} \quad (5.20)$$



**Figure 5.6:** A living frog floating inside a coil with 20T magnetic flux density that was demonstrated by Nobel prize winner Andre Geim. The diamagnetic properties of the water in the frog stabilises its position inside the coil. (Courtesy of Radboud University Nijmegen)

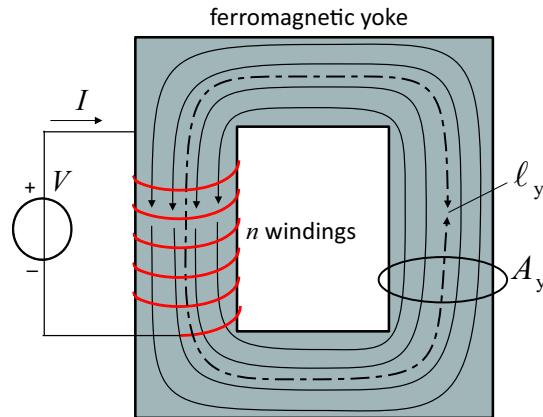
Which is equal to the previously found relations.

Further it also shows clearly the relation between the magnetomotive force  $\mathcal{F}_m = nI$  and the magnetic field strength  $H_w$  of the magnetising field  $\mathbf{H}_w$ .

$$H_w \approx \frac{\mathcal{F}_m}{\ell_i} \quad \Rightarrow \quad \mathcal{F}_m \approx H_w \ell_i \quad (5.21)$$

### 5.1.3.1 Ferromagnetic materials

One method to increase the magnetic flux, without increasing the current, is the use of a material with a higher permeability. Because  $\mu_0$  is the reference, materials other than vacuum have a relative permeability  $\mu_r$  in respect to the permeability in vacuum. This value equals around 1 for non-ferromagnetic materials like air, glass, many metals and plastics but with ferromagnetic materials  $\mu_r$  can reach values of about 100 for iron and nickel to as large as 100,000 for materials like Superpermalloy. Most non-ferromagnetic materials show some very limited effects and are called *paramagnetic* if they show a  $\mu_r$  of just above 1 or *diamagnetic* if they show a  $\mu_r$  of just smaller than 1. The latter have the “strange” property of being repelled by a magnetic field. A famous example is the levitated living frog shown in Figure 5.6. It was demonstrated by the Dutch/Russian Nobel prize winner Andre Konstantinovitsj Geim (1965) at the High Field Magnetic laboratory of the Radboud University in Nijmegen, the Netherlands. The frog is stably



**Figure 5.7:** Magnetic flux generated by a current carrying coil wound around a ferromagnetic yoke is much larger than without the yoke, because of the high magnetic permeability of the ferromagnetic material.

floating inside a very strong magnetic field of about 20T, because of the very small diamagnetic properties of water, which is the main constituent of a frog.

The  $\mu_r$  values of diamagnetic materials are extremely small. For this reason they are still of little practical use in electro magnetic systems and will therefore not be further presented in this chapter. On the other hand real ferromagnetic materials are frequently used for creating strong magnetic fields.

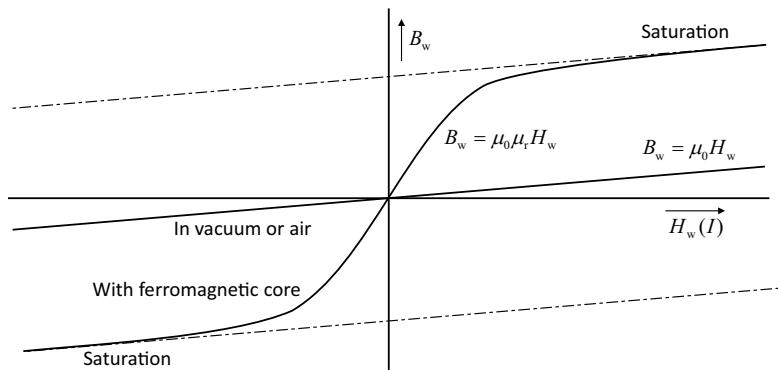
#### 5.1.4 Coil with ferromagnetic yoke

When a yoke of ferromagnetic material is added to a coil, the flux and flux density will increase significantly as the ferromagnetic material determines a low reluctance path for the flux, both inside and outside the coil, as shown in Figure 5.7. By using Hopkinson's law of magnetics the flux in the yoke can be calculated.

The magnetomotive force of the current equals:

$$\mathcal{F}_m = nI \quad (5.22)$$

With a sufficiently high value of  $\mu_r$  the reluctance is mainly determined by the path through the material with the high permeability. This is mostly the case with the usual ferromagnetic materials in mechatronic systems.



**Figure 5.8:** Relation between the flux density  $B_w$  and the magnetic field strength  $H_w$  inside a current carrying coil, with and without a ferromagnetic yoke. Because  $H_w$  is proportional to the current in the coil, an increase in the current will result in an increase of the flux density at a rate determined by the magnetic permeability. With a ferromagnetic yoke, the initially strong effect of  $\mu_r$  will reduce above a certain level of the flux density, due to saturation. At even higher levels the curve will asymptotically continue parallel to the line with  $\mu_0$  only.

With this condition the reluctance becomes:

$$\mathfrak{R} = \frac{\ell_y}{\mu_0 \mu_r A_y} \quad (5.23)$$

with  $\ell_y$  and  $A_y$  being respectively the length and the cross section of the flux path inside the yoke. Note that in spite of the two dimensional case these values are approximated one dimensional where  $A_c$  is assumed orthogonal to the flux and  $\ell_c$  is the average path length.

The flux  $\Phi_y$  in the yoke is equal to the flux  $\Phi_w$  in the coil:

$$\Phi_y = \Phi_w = \frac{\mathcal{F}_m}{\mathfrak{R}} = \frac{A_y \mu_0 \mu_r n I}{\ell_y} \quad (5.24)$$

The corresponding flux density and magnetic field strength inside the coil equal:

$$B_w = \frac{\Phi_y}{A_y} = \frac{\mu_0 \mu_r n I}{\ell_y} \quad \Rightarrow \quad H_w = \frac{B_w}{\mu_0 \mu_r} = \frac{n I}{\ell_y} \quad (5.25)$$

#### 5.1.4.1 Magnetisation curve

From the previous equations it can be concluded, that an increase of the relative permeability  $\mu_r$  of the material of the yoke results in a proportional

increase of the flux and the flux density. In ferromagnetic materials the relative permeability  $\mu_r$  is not constant but reduces at higher levels of flux density. This is visualised in the magnetisation curve, a graphical representation of the flux density  $B_w$  as function of the field strength  $H_w$ . The horizontal axis is also proportional with the magnetomotive force and the current in the magnetising coils, because the field strength is independent of the permeability.

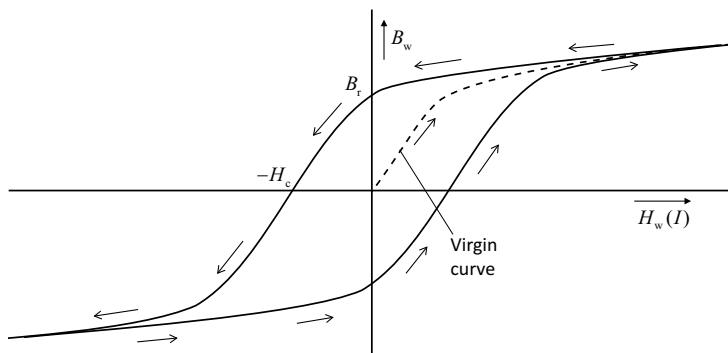
In Figure 5.8 the magnetisation curve of a ferromagnetic material in the configuration of Figure 5.7 is compared with the situation without a ferromagnetic yoke.

The current is increased from zero starting at the origin. This current causes a proportional magnetising field with a magnitude equal to the magnetic field strength  $H_w$ . This also corresponds with an increasing magnetic field of which the magnitude, the flux density, depends on the level of the relative magnetic permeability of the material.

Above a certain value of the flux density, the material becomes saturated. This is caused by small magnetic areas in the materials micro-structure, that are responsible for the relative permeability and have to direct themselves to the magnetic field. These areas are called *Weiss domains* after the French Physicist Pierre-Ernest Weiss (1865 – 1940), who discovered the magnetic orientation of these elementary “building blocks”. A Weiss domain has a typical size of  $10^{-6}$  to  $10^{-8}$  m and contains approximately  $10^6$  to  $10^9$  atoms. The orientation to the external magnetic field takes place inside the Weiss domain and can be imagined as an elastic effect. Initially the domains only partly orient themselves but with higher levels of magnetisation they gradually become completely oriented. As soon as this happens, the beneficial effect to the magnetic field is reduced. When the current is increased beyond this level the flux density increases further proportional to  $\mu_0$  only so parallel to the line corresponding with the situation without a ferromagnetic yoke.

### 5.1.5 Permanent magnets

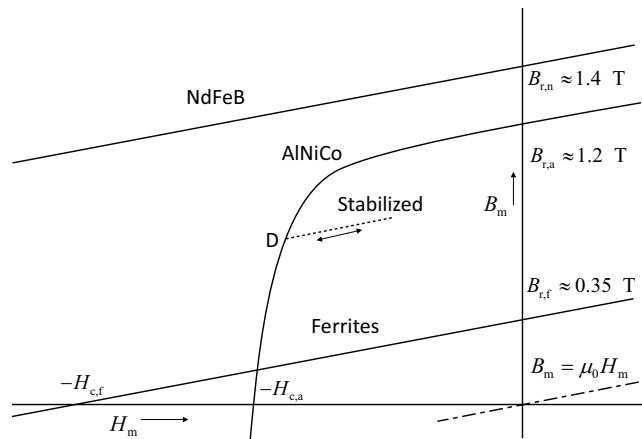
Ferromagnetic materials also show a certain amount of hysteresis. This means that after applying and removing a magnetic field, the material will retain a remnant magnetic property. Like the saturation effect, this is caused in the material itself, where the Weiss domains need some energy to change their magnetic orientation. This can be imagined as a friction effect that limits the possibility for the Weiss domains to change its orientation.



**Figure 5.9:** Hysteresis of a permanent magnet. After the first magnetisation cycle the material will retain some magnetic energy, resulting in a magnetic field that is present, even when the magnetizing field caused by the current is zero. The corresponding flux density  $B_r$  is called the remnant flux density and the negative field strength that would be required to cancel the magnetic field due to the hysteresis, is called the coercive force  $H_c$ .

For many applications this hysteresis is a drawback, as it causes energy loss, due to the need to supply the energy to change the orientation of the Weiss domains. On the other hand however, this special property can help creating magnetic fields, without electrical energy. This is the case in permanent magnets that are utilised in most electromagnetic actuators for precision mechatronics.

A permanent magnet is a ferromagnetic material with a high hysteresis as shown in Figure 5.9 for the same magnetic configuration of Figure 5.7. To illustrate the hysteresis effect, a not yet magnetised material is used at the start of the cycle. In that case the magnetisation by a coil with an increasing current will begin at the origin of the graph and ends until saturation is achieved, like in the previous example. This initial curve on the graph follows the *virgin curve* of the material, because of the not yet magnetised material. When, after saturation, the current is reduced back to zero, the magnetic field strength will by definition also be zero. Due to the hysteresis, however, the magnetic flux density remains positive until a current is applied in the opposite, negative direction. When this negative current is increased, the flux density will follow the upper line of the graph according to the arrows, until the material is saturated in the opposite magnetisation direction. When the current is reversed again, the flux density will follow the lower line until saturation is reached in the other



**Figure 5.10:** Demagnetisation graph of three different kinds of permanent magnets.

The Alnico material exhibits a curved graph with a very high remanent flux density, but needs stabilisation. Magnetic material based on Ferrite (Iron oxides) and the modern very strong material NeFeB have a demagnetisation line parallel to  $B_m = \mu_0 H_m$  and retain their magnetic properties also with occasional strong demagnetisation fields to a level of  $H_c$ .

magnetisation direction again. The virgin curve will only be followed once.

For a permanent magnet only the upper left (or lower right) quadrant of the magnetisation curve is used. Without an external magnetising field, the flux density of a magnetised permanent magnet will be equal to the *remanent flux density*  $B_r$ . The field strength of the internal magnetising field is equal to the external field strength, that would be necessary to cancel the magnetic field. The corresponding magnitude of this field strength is called the *coercive force*  $H_c$ . In search for ever stronger magnets, material science has delivered a wealth of combinations of several, sometimes quite exotic, permanent magnetic materials with very different characteristics. Examples of the *demagnetisation graph* of three different permanent magnet materials are shown in Figure 5.10.

One of the graphs belongs to a magnet material that has been rather popular in the past. It was an alloy of Cobalt, Nickel and Aluminium with some other additives, carrying trade names like Alnico and Ticonal. They showed a large remanent flux density but also a curved graph. This is caused by the Weiss domains that already start to re-orient themselves to the negative magnetic field, giving a relative permeability  $\mu_r > 1$ . Because of the high hysteresis in this material this re-orientation of the Weiss domains becomes

permanent with as a result a reduced remnant flux density. This property is not preferred when these materials are used in electromagnetic actuators that work with current carrying coils. An unintended demagnetisation by the magnetic field of the actuator coil can rather easily lead to a permanent demagnetisation, resulting in a less strong actuator. In the figure this is shown with the dashed line starting at point D. As soon as the magnet is demagnetised until point D, its flux density will follow the dashed line, when the external demagnetising field is reduced again. As a consequence these types of magnets need to be brought to this level on purpose, when external demagnetising fields are expected. This stabilising action is especially necessary with actuators as used in mechatronic positioning systems, and it reduces the potential of these materials to a large extent. It also requires that the magnets are magnetised after mounting in the magnetic circuit, because the high reluctance without a closing magnetic circuit is equivalent to the presence of a high external demagnetising field.

For this reason, ideal permanent magnet materials would need to contain Weiss domains that orient themselves only at magnetic fields with a higher magnitude than those that are present in the magnetic circuit that they are applied in. Fortunately these materials exist, with examples like the frequently used ferrites with iron oxides as base component and the more advanced composites of Samarium-Cobalt and Neodymium-Iron-Boron (Nd-FeB). In these materials the turning of the Weiss domains starts at a field strength larger than  $H_c$ , which means that their relative permeability is equal to one at a field strength below  $H_c$ . As a consequence, the demagnetisation curve runs parallel to  $B_m = \mu_0 H_m$ .

These magnets behave like an air coil without the need to supply electric energy, because the relative permeability of these materials is equal to one. From the fact that an external field strength  $H_c$  is required to compensate the internal field strength, the following relation can be written for the equivalent magnetomotive force of an ideal permanent magnet:

$$\mathcal{F}_m = H_c \ell_m = \frac{B_r \ell_m}{\mu_0} \quad (5.26)$$

In this equation the length  $\ell_m$  of the permanent magnet is equal to the length  $\ell_i$  of the flux path inside the equivalent coil as noted in Equation (5.21).

To illustrate the strength of permanent magnets it is interesting to add some numbers to this equation. For example a permanent magnet with a  $B_r$  of 1 T and a length of 10 mm is equivalent to a coil with a magnetomotive force of  $\mathcal{F}_m = 8 \cdot 10^3$ , which is 80 A in 100 windings. This clearly underlines the

value of permanent magnets as this current would require a lot of electrical power to generate this magnetic field by a small coil only.

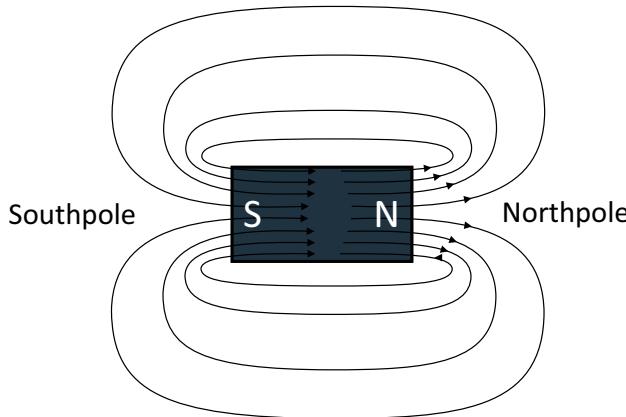
#### 5.1.5.1 Thermal behaviour and Curie temperature

The most important drawback of a permanent magnet, when applied in reliable and predictable mechatronic systems, is the temperature dependency of its demagnetisation graph. This behaviour is caused by the Weiss domains that lose their hysteresis at higher temperatures. The effect is directly related to the Curie temperature  $T_c$ , named after the French physicist and Nobel prize winner Pierre Curie (1859 – 1906), for his investigations on magnetism and piezoelectricity. Above the Curie temperature all Weiss domains become completely disoriented and the magnetic properties are lost. While some permanent magnet materials, like the expensive SmCo alloy, have a rather high Curie temperature with  $\approx 700$  °C, more affordable modern materials like NdFeB alloy have a relatively low Curie temperature of around 300 °C, which limits their use to temperatures below approximately 100 °C.

Even at lower temperatures the magnetic properties of NdFeB alloy decrease with a factor of around  $1 \cdot 10^{-3}$  per °C. Fortunately at these moderate levels the demagnetisation is still reversible and by measuring the temperature, the effect on the magnetism can be compensated.

#### 5.1.6 Creating a magnetic field in an air-gap

In the previous section, the method to create a permanent magnet was explained, starting with a continuous yoke of high-hysteresis ferromagnetic material that was inserted in a coil, like shown in Figure 5.7. After magnetising that yoke, the obtained permanent magnet material would still be rather useless as the magnetic field would remain inside the yoke with a value  $B_r$  of the flux density. When the permanent magnet is to be used to create a magnetic field in an air-gap, it is necessary to create an empty space, an *air-gap* inside the magnet that the magnetic flux has to cross. This can be imagined by cutting the yoke open and aligning it to a straight magnet. The flux generated inside the permanent magnet will exit at the cutting edges, the *pole pieces* to find a return path through the open space as shown in Figure 5.11. It is obvious that this external return path for the flux represents a considerable reluctance and as a consequence the flux will be lower than was the case when the magnet was still enclosed in the



**Figure 5.11:** A piece of permanent magnet material creates an external magnetic field comparable with the field of a current carrying coil, as shown in Figure 5.5.

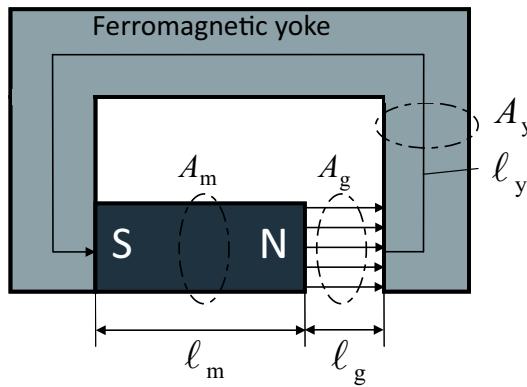
continuous yoke. This means that a de-magnetising field in the opposite direction is created by the external path that forces the flux density down, according to the demagnetisation graph.

This is easier understood by using Hopkinson's law of magnetic fields again, where the magnetomotive force of the permanent magnet creates a flux that is proportional to the sum of the internal and external reluctance. When the external reluctance is zero, like when enclosed in the continuous yoke, the flux is maximum at  $B_r$  and only determined by the internal reluctance. As soon as the magnet is opened with two pole pieces, the external reluctance reduces the flux as mentioned before.

This all implies, that by balancing the reluctance of the external flux path with the internal reluctance of the permanent magnet an optimal use of the permanent magnet material can be achieved with a suitable flux density outside the permanent magnet.

As an example to create an idea about the principle, the magnetic field in an air-gap will be calculated in a roughly approximating way to be able to use it in a Lorentz actuator. This quite standard configuration to create a magnetic field in an air-gap consists of a permanent magnet and a ferromagnetic yoke that is used to concentrate the magnetic field to the air-gap as schematically shown in Figure 5.12.

The equivalent magnetomotive force of the permanent magnet is used to



**Figure 5.12:** Magnetic field in the air-gap caused by a permanent magnet when neglecting flux loss. The reluctance of the magnetic flux path is a combination of the reluctance of the permanent magnet, the air-gap and the ferromagnetic yoke.

calculate the flux in the circuit:

$$\Phi_m = \frac{\mathcal{F}}{\mathfrak{R}_t} = \frac{B_r \ell_m}{\mu_0 \mathfrak{R}_t} \quad (5.27)$$

In this configuration the magnetic reluctance of the complete magnetic path is given by a series of three reluctances that have to be added, the internal reluctance of the permanent magnet itself, the reluctance of the ferromagnetic yoke and the reluctance of the air-gap.

$$\mathfrak{R}_t = \mathfrak{R}_m + \mathfrak{R}_y + \mathfrak{R}_g = \frac{\ell_m}{A_m \mu_0} + \frac{\ell_y}{A_y \mu_0 \mu_r} + \frac{\ell_g}{A_g \mu_0} \quad (5.28)$$

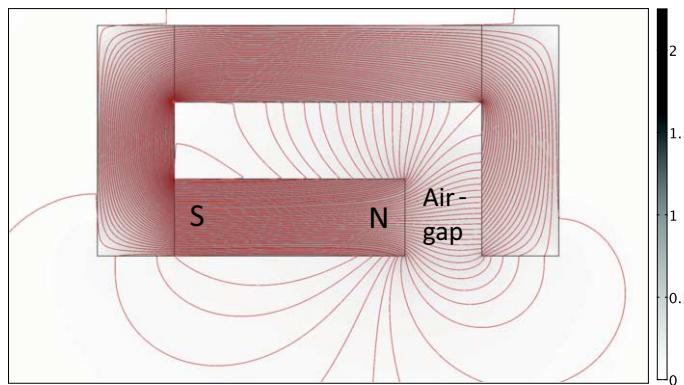
By combining this with Equation (5.27) the following expression for the magnetic flux of the permanent magnet is obtained:

$$\Phi_m = \frac{B_r A_m}{1 + \frac{A_m \ell_y}{A_y \ell_m \mu_r} + \frac{A_m \ell_g}{A_g \ell_m}} \quad (5.29)$$

In many practical situations  $\mu_r \gg \ell_c A_m / A_y \ell_m$  which means that the equation can be reduced to:

$$\Phi_m = \frac{B_r A_m}{1 + \frac{A_m \ell_g}{A_g \ell_m}} \quad (5.30)$$

The flux in the air-gap equals only a part of the total flux of the magnet. The other part is the leakage flux or *stray flux*, that follows a path outside



**Figure 5.13:** Stray flux reduces the useful flux in the air-gap. With FEM modelling software this is clearly made visible in a comparable configuration, as used in Figure 5.12. By counting the flux lines inside and outside the gap the loss factor  $\lambda$  appears to be about 0.45. More than half of the total flux is lost.

the useful area of the air-gap. This stray flux with its corresponding loss factor  $\lambda$  is caused by the fact, that all materials conduct magnetism, even vacuum and air. Magnetic insulation is not possible and this means that the magnetic field will seek the lowest energy situation by using all space available. The magnitude of this effect is shown in Figure 5.13 where in the example configuration a magnet with a  $B_r$  of 1 T creates a flux density in the air-gap of only around 0.3 T and the loss factor  $\lambda$  appears to be approximately 0.45 as can be checked by counting the field lines.

For this reason in this roughly approximating calculation the following relation for the flux  $\Phi_g$  in the air-gap is allowed:

$$\Phi_g = B_g A_g = \lambda \Phi_m = \lambda B_m A_m \quad (5.31)$$

In practice the value for  $\lambda$  will be between 0.25 and 0.75 depending on the configuration. Next to the shown example with  $\lambda \approx 0.45$  two examples with a lower and higher loss factor will be shown as soon as the mathematical analysis is finished.

When Equation (5.30) is combined with Equation (5.31) the magnetic flux density in the air-gap can be calculated:

$$B_g = \frac{\Phi_g}{A_g} = \frac{\lambda \Phi_m}{A_g} = \frac{A_m}{A_g} \frac{\lambda B_r}{1 + \frac{A_m \ell_g}{A_g \ell_m}} = \frac{\lambda B_r}{\frac{A_g}{A_m} + \frac{\ell_g}{\ell_m}} \quad (5.32)$$

This relation reads as follows, by first noting that the flux density in the air-

gap is proportional to the maximum flux density of the permanent magnet. The flux density in the air-gap is reduced by the loss factor  $\lambda$  and further determined by the ratio between the length and cross section of the magnet and the air-gap. The longer the air-gap with respect to the magnet, the less flux density will be achieved and a larger surface of the air-gap in respect to the magnet will also give a lower flux density. In the situation that the cross sections of the magnet and air-gap are equal, like in the example, this equation reduces to:

$$B_g = \frac{\lambda B_r}{1 + \frac{\ell_g}{\ell_m}} \quad (5.33)$$

In the example  $\ell_m = 3\ell_g$  and with  $B_r = 1$  the resulting flux density inside the air-gap would be equal to approximately 0.3 Tesla. More than one decimal is not significant in this very approximating calculation.

### 5.1.6.1 Optimal use of a permanent magnet

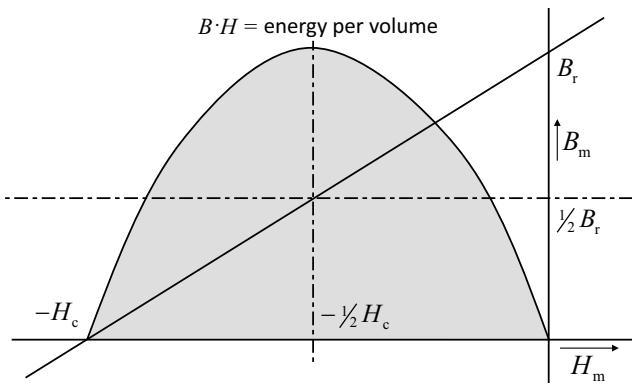
In Figure 5.14 it is shown that a permanent magnet with a straight demagnetisation graph is most efficiently used when the external field strength equals  $H_c/2$  and hence the flux density equals  $B_r/2$ . This is based on the understanding that the specific magnetic energy per unit of volume is proportional to the product of flux density and field strength. This can be checked by using the units:

$$BH = \mu H^2 \implies \frac{A^2}{m^2} \cdot \frac{Vs}{Am} = \frac{\text{energy}}{\text{volume}} \quad (5.34)$$

In the past it was customary, when designing a magnetic circuit with permanent magnets, to work with these values, as it would result in an optimum use of the expensive magnet materials. The discovery of the relatively affordable Neodymium based magnets created the possibility to realise very strong magnetic fields in less optimal designs from a magnet material usage point of view. These configurations are applied in extreme high performance positioning systems where the cost of the magnets is less important. In the following part the more optimal configurations will be presented, as they are a good starting point for gaining further knowledge in this field.

### 5.1.6.2 Flat magnets to reduce stray flux

In order to reduce the stray flux and optimise the use of expensive magnet materials it is often preferred to use a different configuration than the



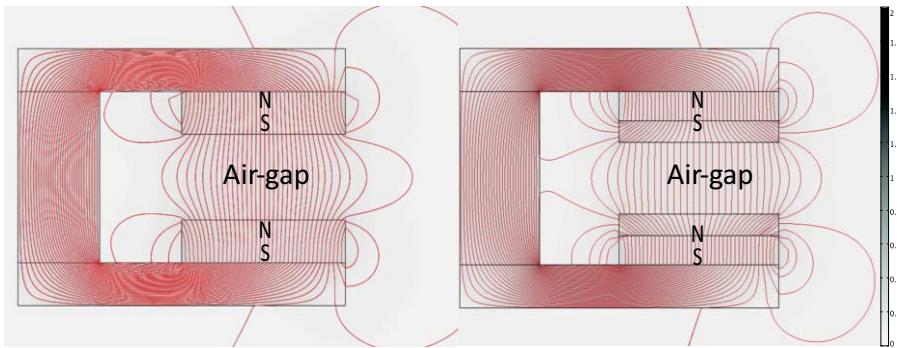
**Figure 5.14:** The magnetic energy of  $B \cdot H$  of a permanent magnet with a straight demagnetisation graph is maximum at half the value of  $B_r$  and  $H_c$ . When the magnetic circuit is dimensioned such that the permanent magnet is utilised at its maximum magnetic energy, the volume and cost of the magnetic material is minimised.

one shown in Figure 5.12 by choosing the cross section of the magnets large in respect to their length. It also appears to be useful to not include a ferromagnetic part between the magnet and the air-gap as the ferromagnetic part would determine a low reluctance path to the sides of the air-gap even for the flux coming from the inner part of the magnet. This is shown in Figure 5.15, where at the right side the effect of ferromagnetic pole pieces is shown and at the left side the optimal configuration with permanent magnets directly adjacent to the air-gap.

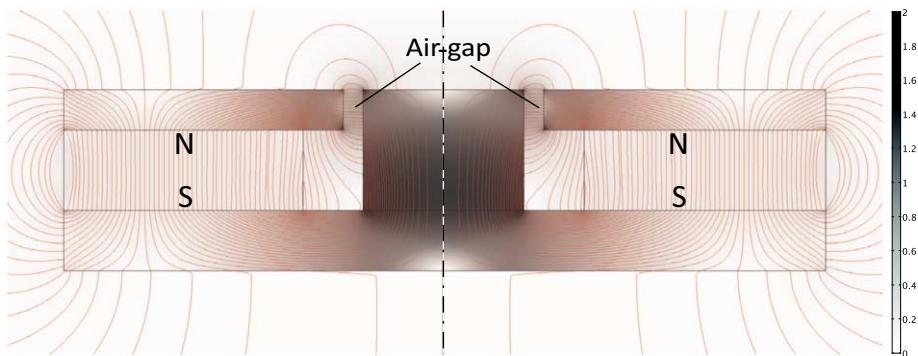
In this configuration the path length though air is often equal to the path length through the magnets and the flux density in the air-gap using Equation (5.33) becomes:

$$B_g = \frac{\lambda B_r}{2} \quad (5.35)$$

When applying magnets with a  $B_r$  of 1 T this second example shows a flux density in the air-gap of approximately 0.4 T. The flux density in the permanent magnet can be determined by using Equation (5.31) with a loss factor  $\lambda = 0.75$ , as found by counting the magnetic field lines. Because  $A_m = A_g$  this calculation results in a flux density of the magnet  $B_m$  of approximately  $B_r/2$ , which means that the permanent magnet is used at its optimal energy point.



**Figure 5.15:** With flat permanent magnets, directly interfacing the air-gap, the stray flux is kept at a minimum. Often ferromagnetic pole pieces are suggested to spread the magnetic field more uniform, however, as shown in the right figure, this pole piece significantly increases the stray flux, due to the low reluctance path to the edges of the magnet. Even with a smaller air-gap the loss factor  $\lambda$  reduces from approximately 0.75 to 0.6 as can be checked by counting the flux lines.



**Figure 5.16:** Concentration of flux by ferromagnetic pole pieces. This rotation symmetric structure with a large but inexpensive magnet, as commonly applied in loudspeakers, creates a strong magnetic field directed in the radial direction in the air-gap. A large part of the magnetic flux is however lost outside the air-gap.

### 5.1.6.3 Low cost loudspeaker magnet configuration

When a very high flux density has to be created in a relatively small air-gap, it is possible to concentrate the flux by using ferromagnetic pole pieces that have a smaller surface at the air-gap than at the surface of the magnet. Even though this creates an increased loss it is often used with very inexpensive

magnets like the ferrites that are usually applied in loudspeakers. The effect on the flux is shown in Figure 5.16.

With Equation (5.32) this concentrating effect can be calculated, although due to the high amount of stray flux ( $\lambda < 0.3$ ), the magnet needs to be considerably larger than would be the case with the flat magnets of Figure 5.15. Still in this example an inexpensive magnet with a  $B_r$  of 0,4 T creates a flux density in the air-gap of approximately 0.7 T , which is fully acceptable when low cost is the primary specification and size is no issue.

In the next section the magnetic field in the air-gap is used to create a Lorentz actuator by inserting a current conducting wire in the magnetic field.

## 5.2 Lorentz actuator

Lorentz actuators are predominantly applied in high precision positioning systems because of their inherent low mechanical stiffness between the stationary and the moving part. Also the linear relation between current and force combined with the favourable dynamic properties are important factors. The low stiffness reduces the amount of external motion that is transferred from the support structure through the actuator to the moving part (transmissibility!). These movements can be caused by vibrations of surrounding machines but also by the reaction forces of the actuator itself, exciting resonances in the support structure. As will be shown later, Lorentz actuators have also some drawbacks like the relatively modest force to current ratio which limits the maximum acceleration levels and the achievable range of motion or “stroke”. As the name implies, the Lorentz actuator is based on the Lorentz force only.

### 5.2.1 Lorentz force

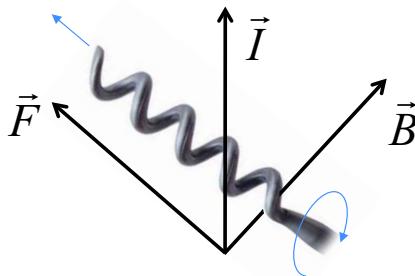
The Dutch physicist and Nobel prize winner Hendrik Antoon Lorentz (1853 – 1928) formulated the Lorentz force as a completion to the Maxwell equations. The law of Faraday describes the effect of a changing magnetic field on electrical charges hence generating electricity from kinetic energy. Based on energy conservation laws creating electrical energy from motion is fully complementary to creating motion energy from electrical energy so the laws of Lorentz and Faraday are strongly related.

In vectorial notation the formulation of Lorentz describes the force on a moving charged particle as:

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \quad (5.36)$$

with  $\mathbf{v}$  [m/s] equals the instantaneous velocity of the particle. The first part of the Equation  $q\mathbf{E}$  is the electrostatic force and the second part is the electromagnetic force. This second term is used in electromagnetic actuators. Next to the force on a moving particle it equally represents the force on a current flowing through a wire with length  $\ell_w$  [m], inserted in the magnetic field. For this situation the moving charge equals the current times the length,  $q\mathbf{v} = \ell_w \mathbf{I}$ , and with this relation the electromagnetic Lorentz force is equal to:

$$\mathbf{F} = \ell_w \mathbf{I} \times \mathbf{B} \quad (5.37)$$



**Figure 5.17:** Determining the direction of the Lorentz force with the corkscrew rule.

When the corkscrew is rotated right handed, from the direction of the positive current to the direction of the magnetic field (arrow), the movement of the point of the corkscrew determines the direction of the force.

For the magnetic force on a wire at an angle  $\alpha$  relative to the direction of a magnetic field with flux density  $B$ , carrying a current  $I$ , this relation leads to the scalar notation of the Lorentz force of electromagnetic actuators of which the magnitude is given by:

$$F = BI\ell_w \sin \alpha \quad (5.38)$$

The direction of this force is orthogonal to the plane that is determined by the direction of the magnetic field and the current, due to the “cross product” in the vectorial Lorentz equation. This rule can be remembered as the *right hand* or *corkscrew* rule that states that the positive force direction is found when rotating a corkscrew from the positive current direction onto the direction of the magnetic field as shown in Figure 5.17. Of course for a real mechanical engineer any normal right turning screw will also suit the purpose, but the corkscrew is more easy to remember.

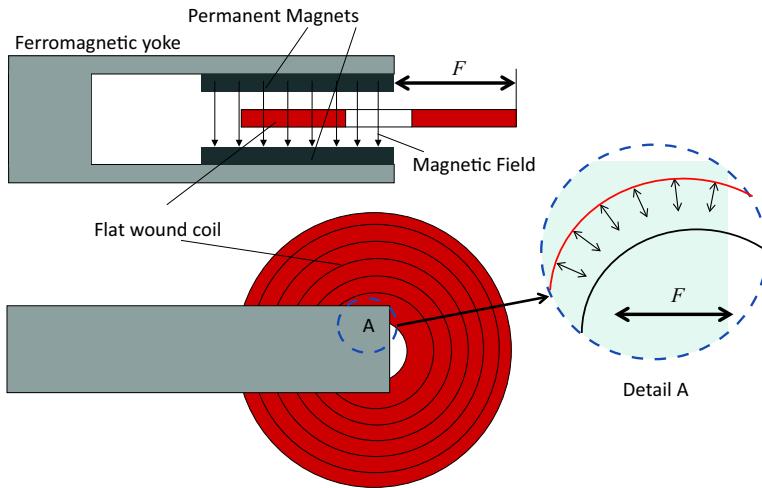
In most practical cases the Lorentz force must be maximised which means that  $\sin \alpha$  is kept as much as possible equal to one. This means that the simplified equation becomes equal to:

$$F = BI\ell_w \quad (5.39)$$

And with multiple windings the Lorentz force becomes:

$$F = BIn\ell_w = BI\ell_{w,t} \quad (5.40)$$

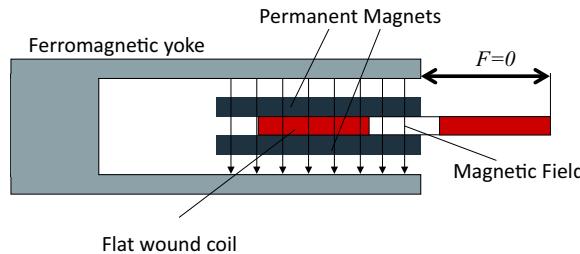
where  $\ell_{w,t}$  equals the total length of the wire inserted in the magnetic field. This equation is widely used as a general equation for linear actuators based



**Figure 5.18:** Basic flat type Lorentz actuator. The Force  $F$  is proportional to the current and the amount of winding length  $\ell$  of the wire in the coil inserted in the magnetic field  $B$ . Due to the large amount of coil outside the air-gap this is not a very efficient configuration. Detail A shows the forces acting on a wire segment at the centre of the coil, clearly indicating the inefficiency.

on the Lorentz principle. It clearly shows that the force only depends on the current, the total wire length and the magnetic flux density. Because there is no direct relation with the position, the actuator ideally has zero stiffness. This is the main reasons why a Lorentz actuator is preferred in precision positioning systems as it avoids transmissibility of vibrations from the stationary part to the moving part. Later it will be shown that this is the ideal situation. In reality at higher current levels some stiffness is observed due to non-linearity and position dependency of the  $B\ell$  value.

Figure 5.18 shows a basic flat type Lorentz actuator configuration for illustrating the principle only as it is not the best design possible. It uses the flat high efficiency permanent magnet configuration as described in the previous section and the coil is flat wound and inserted partly in the magnetic air-gap. The force of this actuator can in principle be calculated using Equation (5.40). The permanent magnet flux is directed perpendicular to the current. When the curvature of the wires is neglected, the force equals approximately  $BI\ell_{w,t}$ , where  $\ell_{w,t}$  is the part of the total length of the wire inside the air-gap. The inner windings contribute little to the force, because of the curvature. This is shown in Detail A that indicates the spread force contributions along a wire segment at the centre. Only the component of



**Figure 5.19:** Non operating actuator with mechanically connected magnets and coils. The force is created between the magnetic flux and the current and not between the magnet-coil combination and the iron part of the system.

the force pointing to the right contributes to the total force.

A drawback of calculating the force with the simplified equation is however that it sometimes leads to mistakes. To illustrate this, a real life example of a non-operating Lorentz actuator is shown in Figure 5.19. Originally the idea was meant to increase the stroke by extending the iron parts and move the magnets and the coil together. Unfortunately, this idea does not work at all, even though the current carrying coil is inserted in an area with a significant magnetic field. The reason of the lack of observed force is, that the force acts between the current and the permanent magnet flux. As a consequence, no external force will occur when the coil and the magnets are mechanically connected. To avoid these errors, it is preferable to work with another more generally suitable relation that is based on the change of flux over the windings as function of the movement of the coil relative to the magnetic field. When one side of a closed winding is inserted in a magnetic field and it moves with a value  $dx$  relative to the permanent magnetic system in the direction of the force, the flux  $\Phi_w$  inside the winding will change according to the flux density and the length  $\ell_w$  of the inserted part of the winding:

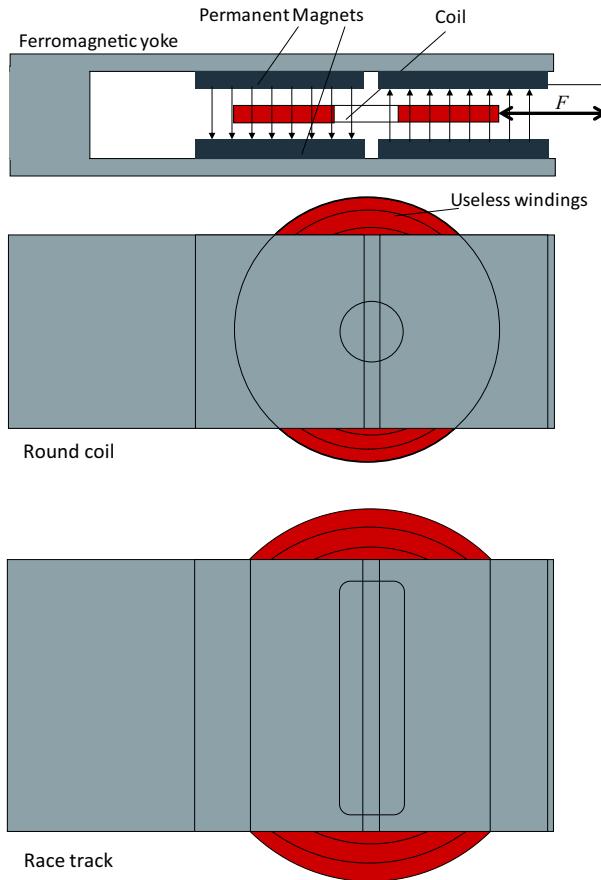
$$d\Phi_w = B \ell_w dx \implies \frac{d\Phi_w}{dx} = B \ell_w \quad (5.41)$$

The Lorentz force for each winding can now also be written as:

$$F_w = I \frac{d\Phi_w}{dx} \quad (5.42)$$

And for multiple windings the relation becomes:

$$F = nI \frac{d\Phi_w}{dx} \quad (5.43)$$



**Figure 5.20:** Flat type Lorentz actuator with improved Force to current ratio. By increasing the width and adding two additional magnets like shown in the upper drawing, the force is doubled at the same power loss in the windings. Using an oval shape of the coil like a “race track” as shown in the lower drawing, will improve the force to power ratio even more.

Using this equation it is obvious that the actuator of Figure 5.19 is not working, because with that configuration there is no change of the flux through the coil as function of the displacement ( $d\Phi_w/dx = 0$ ).

## 5.2.2 Improving the force of a Lorentz actuator

In the previous section it was demonstrated that the windings of the Lorentz actuator of Figure 5.18 are not very well utilised. While the resistive power

loss ( $P_1 = I^2 R$ ) of an actuator is directly proportional with the length of the windings, the actuator should be designed with a maximum coverage of the coil by magnetic flux density and a maximum contribution of the force of each part of the windings in the right direction. A first improvement can be achieved by using both sides of the coil. At first sight it might be useful, to extend the permanent magnet, in order to cover the entire coil. Unfortunately however all forces would direct to the centre of the coil and cancel each other out. This corresponds with the fact that in that situation  $d\Phi_w/dx = 0$ . This problem can be solved by reversing the magnetic field of the right side of the coil where the current runs in the opposite direction of the current in the left side. By adding two permanent magnets, magnetised in the opposite direction of the first set of magnets and widening the magnet system, as shown in the upper drawing in Figure 5.20 the coil is almost completely used. As a result the useful force is doubled in respect to the first example, with the same power loss in the coil.

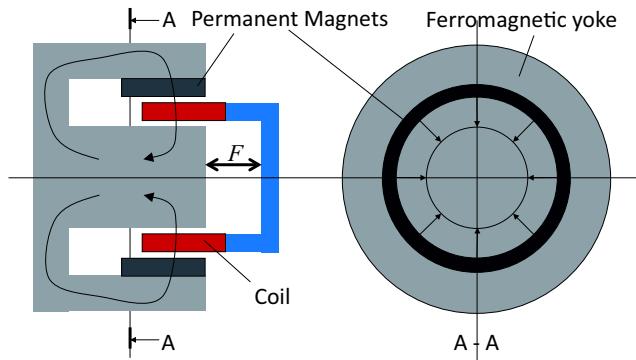
A further improvement can be achieved by changing the coil into an oval shape as shown in the lower drawing in Figure 5.20. This is known as a *race track coil* and this configuration has become the “de facto” standard in flat Lorentz actuators for fast and high precision positioning systems, because of the optimal use of the windings.

### 5.2.3 The moving-coil loudspeaker actuator

The most widely applied version of the Lorentz actuator is the moving-coil loudspeaker actuator. This type is a fully rotation symmetric configuration with a round coil that is inserted in a magnetic field that is directed radial towards the centre of the structure, like shown in Figure 5.16. Even though the magnetic field in the air-gap is three dimensional the orientation is orthogonal to the current of the windings. For this reason the approximating scalar expression for the force can be used also in this example. A version with permanent magnets directly adjacent to the air-gap with the coil is shown in Figure 5.21. In this configuration the coil is completely surrounded by the permanent magnet, resulting in a maximum efficiency.

### 5.2.4 Position dependency of the Lorentz force

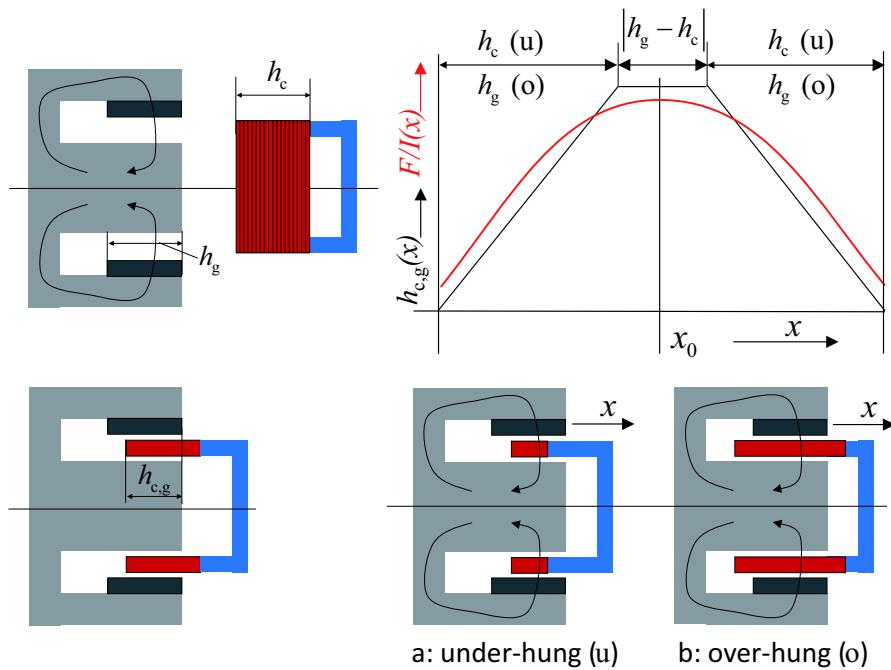
A Lorentz actuator, like the moving-coil loudspeaker type, shows a force to current ratio ( $F/I$ ) that is dependent on the position of the coil in the air-gap. This effect is shown in the red line of the graph in Figure 5.22



**Figure 5.21:** moving-coil loudspeaker motor with permanent magnets directly adjacent to the air-gap. In the middle position the coil is fully inserted in the magnetic field giving a maximum force to power ratio.

and is caused by the fact, that off the centre, the coil is not completely surrounded by the same magnitude of the magnetic field. When used in a closed loop positioning system this implies a change in the control loop amplification (gain) and a non-linear positive or negative stiffness depending on the current level and the position relative to the centre position  $x_0$ .

This can be explained as follows. Assume the coil carries a current with a direction that corresponds to a force in the positive  $x$  direction. When the coil is pushed by an external force  $F_x$  in the negative  $x$  direction from outside into the air-gap, at first the force needed to move the coil increases. This is due to the increased part of the height  $h_{c,g}$  of the coil inside the air-gap and continues until at  $x_0$  the maximum force is reached. As a consequence, this trajectory showed a positive stiffness,  $dF/dx \geq 0$ . When the coil is moved further in the negative  $x$  direction the force decreases again, which corresponds with a negative stiffness,  $dF/dx \leq 0$ .



**Figure 5.22:** A Lorentz actuator has a limited stroke, determined by the dimensions of the air-gap and the coil. When the coil is only partly inserted with an effective height  $h_{c,g}$  that is inserted in the air-gap, the Force to current ratio ( $F/I$ ) is reduced. By choosing different values for the height  $h_c$  of the coil and the height  $h_g$  of the air-gap, the Force to current ratio as function of the position  $x$  can be made more constant over a certain range of  $x$ . The gradual decrease of the magnetic field (stray flux) at the edges of the air-gap softens the transitions.

### 5.2.4.1 Over-hung and under-hung coil

In order to reduce the position dependency of a Lorentz actuator, in practice the height  $h_g$  of the air-gap is chosen different from the height  $h_c$  of the coil. When the coil height exceeds, hangs over, the height of the gap, this configuration is called an *over-hung* Lorentz actuator.

Its advantage is the optimal use of most of the permanent magnetic flux while also the position dependency is more evenly smoothed'out. The drawback of an over-hung actuator is the large number of coil windings, that is not utilised effectively. In the situation, where the coil is the moving part, this means that the moving mass is larger than without an over-hung situation. It is a typical choice for loudspeakers where the cost of the magnet

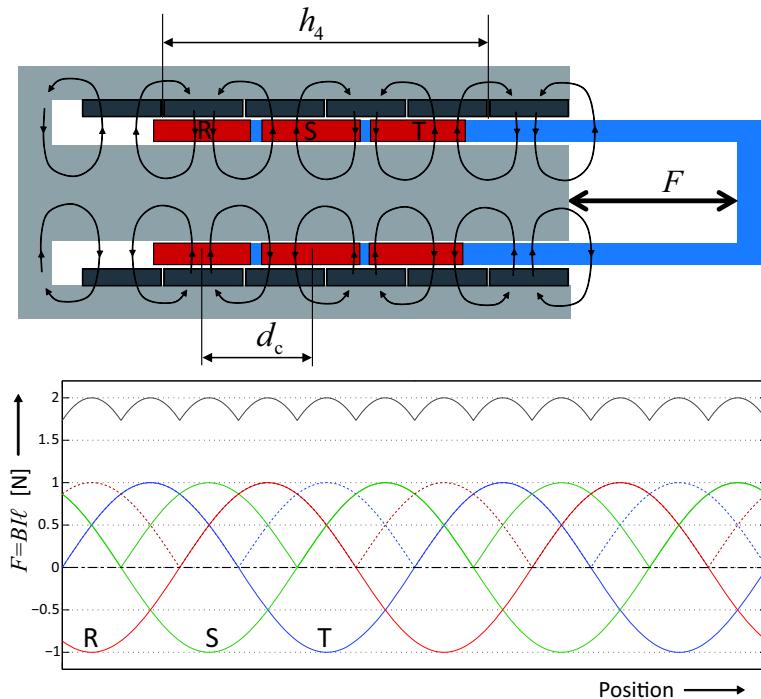
outweighs the value of a high efficiency of the loudspeaker. The amount of overhang is chosen depending on the allowed non-linearity related distortion. For non-critical applications like loudspeakers for cars, "Public-address" in stadiums and amplification of instruments in pop-music, no overhang is applied, giving a maximum efficiency at increased distortion. For high-end sound reproduction, often a large overhang is chosen to reduce distortion at a sacrifice on efficiency. This is the reason why amplifiers for high-end home equipment need to have a relatively high output power.

The configuration where the height of the gap exceeds the height of the coil is called an *under-hung* Lorentz actuator. The benefits and drawbacks are just reversed. It makes best use of the coil with a resulting reduced mass but at a relative high cost of the permanent magnetic part. The position dependency of the force to current ratio is better than in the over-hung configuration, when the coil is still completely inside the air-gap, but it worsens more rapidly as soon as the coil reaches the outer range of the air-gap, due to its small size. An under-hung Lorentz actuator is chosen mainly in loudspeakers, when a long stroke is not needed and low moving mass and distortion is the primary goal. This is especially the case for mid-to high-frequency loudspeakers, the squawkers and tweeters.

### 5.2.5 Electronic commutation

The only method to increase the range of a Lorentz actuator, while preserving a constant force to current ratio, is by combining several coils and magnets into a *electronically commutated* actuator. The principle is shown in Figure 5.23. A set of an even number of alternately magnetised permanent magnets creates an alternating field in the air-gap. The coil is divided in three equal separate sections. In Figure 5.22 it was shown that the force factor  $B\ell$  of each coil segment is dependent on the position. By choosing an optimal height relative to the permanent magnets, the force factor of each coil segment becomes an almost ideal sinusoidal function of the position. When the centres of the segments have a distance  $d_c$  equal to one third of the total height  $h_4$  of four magnets the spatial sinusoidal forces have a spatial phase difference of  $120^\circ$ .

This is a *three-phase* actuator configuration. The relation between the sinusoidal functions is completely comparable to three-phase power distribution networks where three temporal sinusoidal voltages, called R, S and T with a phase difference of  $120^\circ$ , are used to transport electrical energy. One reason to use this method is the fact that the three sinusoidal functions add to zero



**Figure 5.23:** The range of a Lorentz actuator can be extended by electronic commutation, using more coils and magnets. Each coil segment has a force factor that changes sinusoidal as function of the position with a  $120^\circ$  spatial phase difference with the other coil segments. By changing the direction of the current at the zero force positions of each coil, an almost constant force is obtained. The values are modelled for a current per coil of 1 A and a value of  $(B\ell)_{\max} = 1$ .

at any position in the period, as can be proven by applying trigonometry on the following function.

$$\begin{aligned} \sin(x) + \sin\left(x - \frac{2}{3}\pi\right) + \sin\left(x - \frac{4}{3}\pi\right) &= \\ \sin(x) + \sin(x)\cos\left(\frac{2}{3}\pi\right) - \sin\left(\frac{2}{3}\pi\right)\cos(x) + \\ + \sin(x)\cos\left(\frac{4}{3}\pi\right) - \sin\left(\frac{4}{3}\pi\right)\cos(x) &= 0 \end{aligned} \quad (5.44)$$

With power distribution this means that the average voltage of three wires carrying these voltages is zero at any time. This prevents electromagnetic

radiation from the power lines. Another important value is the possibility to directly drive rotating AC inductance motors for traction and machining centres. In mechatronic positioning systems it appears to be also very useful in long range linear motors.

When all coil segments in the configuration of Figure 5.23 would get the same DC current the resulting force would be zero. This might seem useless, but when the current of each coil is reversed at the exact position ,where the force in that segment is zero, the forces of all three segments add to an almost constant value of two times the maximum value of  $B\ell$  per coil section, with a variation as indicated in the figure.

In rotating DC motors this current commutation process is mostly done by means of mechanical sliding contacts, although presently this commutation is increasingly achieved electronically in order to avoid sparking at the commutation points and wear of the sliding contacts.

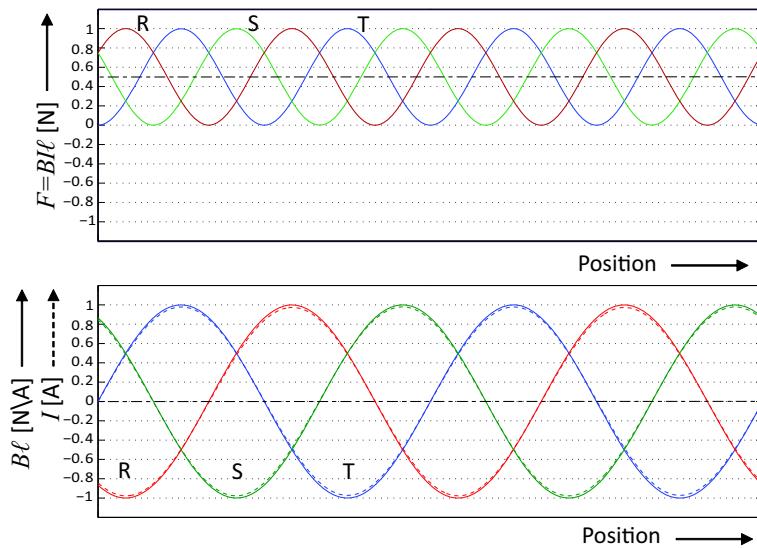
The application of electronic commutation can also help solving the problem of the not perfect continuous force by exchanging the hard-switching by a more continuous current change.

### 5.2.5.1 three-phase electronic control

By using a special three-phase amplifier as will be presented in Section 6.3 of Chapter 6 the current to each coil segment can be controlled in such a way that the magnitude changes with the same sinusoidal function of the position as the force factor. The resulting force then becomes equal to the multiplication of the amplitude  $\hat{I}$  of the current, the amplitude of the force factor  $(B\ell)_{\max}$  and the sine function squared. Figure 5.24 shows the signals and forces when the current and  $(B\ell)_{\max}$ -value are equal to the hard-switched example of Figure 5.23.

The squared sinuses produce a three-phase combination at double the spatial frequency of the original sine wave and these sinuses add to three times their average value. For the example with 1 A current and a maximum  $(B\ell)_{\max}$  value of one, the resulting force would be constant at a level of  $3 \cdot 0.5 = 1.5$ .

Even though this force level is slightly below the level with the hard-switching commutation of Figure 5.23, the method of three-phase electronic commutation is preferred, because of the constant force and the absence of fast transients in the currents and forces that might introduce electromagnetic interference and excite uncontrollable eigenmodes in the mechanism.



**Figure 5.24:** three-phase commutation with a sinusoidal control of the currents in each coil segment in phase with their ( $B\ell$ ) factor results in a force per segment with a spatial frequency that is double the original spatial frequency of the coils around an average value that equals half the  $(BI\ell)_{\max}$ -value. The resulting total force of the three coil segments is the sum of the average values of the force in each segment and is independent of the position.

### 5.2.6 Figure of merit of a Lorentz actuator

Before presenting the non-linear behaviour of electromagnetic actuators a short story is given on how at Philips Electronics, the design of a Lorentz actuator was continuously improved by working with a *figure of merit*.

During the early years of developing actuators for the CD drive like presented in Chapter 1 many different design concepts were investigated. To arrive at a constant current to force ratio [N/A] over a stroke of about one millimetre one could use a under-hung “short coil - long magnet” combination or an over-hung “long coil - short magnet”. Also designs using the coil as mover and designs with moving magnets were made. When actuating in only one or two directions the remaining degrees of freedom must be restrained by guiding elements. Both sliding elements and flexible designs are possible. Similarly one could use translation or rotation as primary kinematic solution for the motion. Looking at all these options, which have all been tried, it became clear that decision making was not easy. To settle this a criterion

was derived that was successfully used for that purpose, the figure of merit. The basic function of the actuator is to generate a certain acceleration of the optical element. Driving with higher currents can deliver more acceleration but the penalty is a higher dissipation. The final “Figure of Merit” was called: “G’s per square root of Watt’s”. How many G’s one would get per watt of power. The higher the figure the better. If the figure of merit is called  $Q_m$  the relation is:

$$Q_m = \frac{G}{\sqrt{\text{Watt}}} \quad (5.45)$$

Although a bit unconventional, unscientific and at least not according to SI units this number proved to be quite suitable and has been in use for about 10 years. When analysing the figure, using the actuator design aspects, the reason becomes clear. As a first step, the figure of merit is written as follows according to SI rules:

$$Q_m = \frac{F}{\sqrt{P}} = \frac{m}{\sqrt{P}} = \frac{F}{m\sqrt{P}} \quad (5.46)$$

With  $a$  as acceleration and  $P$  as power. In such a mechanism this power almost completely consists of the resistive power loss ( $P = P_1$ ), as the movements that are made are very small. This relation already makes clear that a small mass is favourable. A large force at low dissipation also leads to a higher score. One step further  $F$  can be eliminated, using the simple Lorentz actuator formula  $F = BI\ell_{w,t}$  with  $\ell_{w,t}$  being the total length of the wires inside the magnetic field of the permanent magnet.

$$Q_m = \frac{BI\ell_{w,t}}{m\sqrt{P_1}} \quad (5.47)$$

The second step is related to geometric aspects. When the part of the length of each winding in the air-gap (active part) is called  $\ell_{w,a}$  and the length of the part of each winding outside the air-gap (passive part)  $\ell_{w,p}$  and the number of windings is  $n$ , it is allowed to say:

$$Q_m = \frac{BIN\ell_{w,a}}{m\sqrt{P_1}} \quad (5.48)$$

Next the dissipation can be eliminated by using  $P_1 = I^2R$ . Upon substitution it is shown, that the current is present both in the nominator and denominator so it cancels out:

$$Q_m = \frac{Bn\ell_{w,a}}{m\sqrt{R}} \quad (5.49)$$

The resistance  $R$  is the total resistance, including the active part and the passive part of each winding. With  $n$  windings of a resistivity  $\rho_r$  and a cross section per wire  $A_w$ , the resistance of the coil becomes:

$$R = \frac{n \rho_r (\ell_{w,a} + \ell_{w,p})}{A_w} \quad (5.50)$$

With  $\gamma$  being the fill factor of the windings and  $A_{c,w}$  the cross section of the coil windings ( $b_c h_c$ ), the cross section per wire becomes:

$$A_w = \frac{\gamma A_{c,w}}{n} \quad (5.51)$$

With this value the resistance equals:

$$R = \frac{n^2 \rho_r (\ell_{w,a} + \ell_{w,p})}{\gamma A_{c,w}} \quad (5.52)$$

The active part of the volume of the coil within the magnetic field  $V_{c,a}$  is equal to  $A_{c,w} \ell_{w,a}$ . When substituted together with Equation (5.52) in Equation (5.49), this leads with a bit algebra to the following expression for the figure of merit:

$$Q_m = \frac{B}{m} \sqrt{\frac{\gamma V_{c,a}}{\rho_r}} \sqrt{\frac{\ell_{w,a}}{(\ell_{w,a} + \ell_{w,p})}} \quad (5.53)$$

It is demonstrated, that the actuator figure, “G’s per square root of Watt’s” has been changed into clear understandable and far more logical engineering choices. By maximising  $B, \gamma, V_{c,a}$  and minimising  $m, \rho_r$  and  $\ell_{w,p}$  relative to  $\ell_{w,a}$ , it became possible to realise an optimal design at such a low price, that this part of the CD player changed from a costly part as in the first CD players to the mass produced cheap commodity it is today.

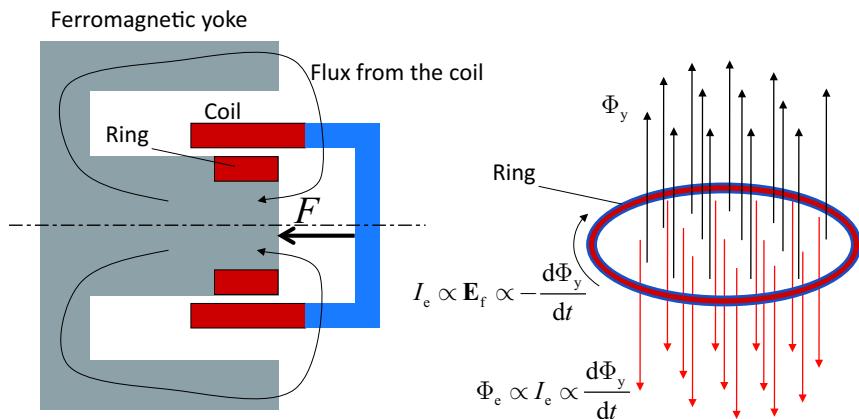
## 5.3 Reluctance actuator

The force of an electromagnetic actuator is not always linearly related to the current like in an ideal Lorentz actuator. In general, at increased current levels, the force will change more disproportional due to different causes. First of all the magnetic field of the coil adds to the magnetic field of the permanent magnets. As a result the iron parts in the magnetic assembly can get saturated, with as a consequence an increased reluctance with a corresponding reduction of the magnetic flux by the permanent magnet. This means that at higher current levels the force will be smaller than expected from the linear relation  $F = BI\ell_w$ . An other source of non-linearity is the direct interaction of the iron part and the coil. The coil will act as an electromagnet, attracting the iron in a direction that is independent of the direction of the current. This attractive force is called the *reluctance force*, because it is related to the change of reluctance when the coil approaches the iron, and its magnitude is proportional to the current squared.

The reluctance force is used to create many actuators as will be presented later in this chapter but its impact on linearity is a drawback when applied in a precision positioning system. As a first step in the treatise of the reluctance force, it will be examined as an unwanted effect in a Lorentz actuator with possible means to reduce it. After this analysis, the analytical force equations of electromagnetic actuators of any type will be derived from the law of conservation of energy. It will be followed by a presentation of the variable reluctance actuator and the permanent magnet biased reluctance actuator.

### 5.3.1 Reluctance force in Lorentz actuator

In precision positioning systems, the Lorentz actuator is chosen mainly in situations where no transmissibility due to motor stiffness is allowed. As was demonstrated, Lorentz actuators are not completely free of stiffness. The first cause for residual stiffness is based on the relation between force and position and was discussed in the previous section. It represented a non-linear stiffness, that can be kept at a low level by limitation of the movement at the centre or by using electrical commutation. In case of a short stroke actuator, the reluctance force can cause another stiffness factor. Figure 5.25 shows the induced flux by the current in the coil with an actuator where the coil is located slightly outside the optimal position in the middle of the air-gap. The permanent magnets are not shown, because the reluctance



**Figure 5.25:** The reluctance force in a Lorentz actuator is a force that acts independent of the permanent magnet flux. It is caused by the attraction of the iron part by the magnetic field of the coil. For high frequencies this force can be reduced by a conductive ring, that counteracts the change of the magnetic flux in the iron part, as shown at the right.

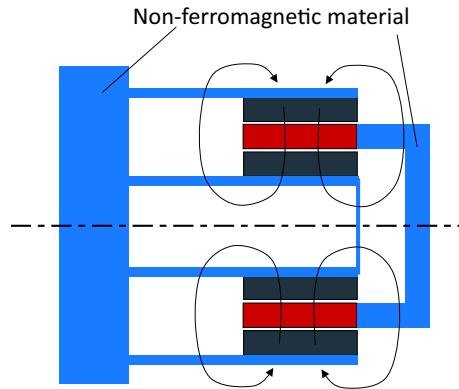
force is only determined by the ferromagnetic part.

It is clear, that the reluctance of the flux caused by the current in the coil is minimal when the coil is located completely inside the ferromagnetic part. The reluctance force is related to a maximum in the magnetic flux as function of the position. This means that the reluctance force on the coil is always in the direction of the ferromagnetic part.

One way of reducing this reluctance force would be to extend the ferromagnetic part to the right to beyond the magnets but that would increase the mass and requires the coil holder to be longer.

### 5.3.1.1 Eddy-current ring

A second method of reducing the reluctance force is by using a conductive ring, connected to the ferromagnetic part as shown in Figure 5.25. The effectiveness of this method is based on the variability of the flux from the coil. The change of the flux  $\Phi_y$  inside the ring, by the current in the coil, induces an inner Electric field  $E_f$  over the ring according to Faraday's law. Because the ring is closed, this electric field causes the electrons to move, resulting in a current  $I_e$  in the ring in the direction of the electric field, as explained in Chapter 2 on electricity. This current is also called an *eddy-current*, because in principle it is induced in any conductive material



**Figure 5.26:** Lorentz actuator without a ferromagnetic part to cancel the reluctance force. The return path of the permanent magnet flux goes only through air.

inserted in a changing magnetic field and it behaves like circular running currents inside the material like the “eddy” currents or swirl in a fast flowing river. In its turn, this induced eddy-current in the ring causes a magnetic field  $\Phi_e$  according to Ampère’s law in the opposite direction of the magnetic flux of the coil. This means that the change of the magnetic flux of the coil inside the ring is suppressed.

The total flux inside the ring equals the integral of the change of the flux over time. As the change of flux increases with frequency, the reducing effect of the ring on the flux, and correspondingly also on the reluctance force, is most effective at higher frequencies and it is completely absent at steady state currents. Because of the limitation to higher frequencies this method is frequently applied in loudspeakers. For precision mechatronic systems it is however not suitable, as those require a reliable operation, also with non-varying, steady state forces.

### 5.3.1.2 Ironless stator

For those more critical applications where the linear performance is far more important than the cost of a bit more magnet material, one can decide to entirely leave the ferromagnetic part away. Figure 5.26 shows such a configuration, where the magnets are connected by means of non-ferromagnetic material. A second ring of magnets has been added to compensate for the higher reluctance at the inside of the magnetic system. The increase of the reluctance for the permanent magnet flux at the outside is less dramatic for

the same reason as was presented with the air coil from Figure 5.4, where the reluctance of the path outside the coil could be neglected in respect to the reluctance inside the coil.

This configuration is used in the wafer stage that is presented in Chapter 9 on wafer scanners.

### 5.3.2 Analytical derivation of the reluctance force

As will be explained with the Reluctance actuator in the next section, Equation (5.43) is not valid under all circumstances as it is limited to the situation where  $d\Phi_w/dx$  is only depending on the permanent magnet and not on the current. This means that it is necessary to use a different approach to determine the total force that acts on an electromagnetic actuator including the reluctance force. A suitable method is based on the law of conservation of energy in a closed system.

In an electromagnetic actuator, electrical energy is converted in three different kinds of energy:

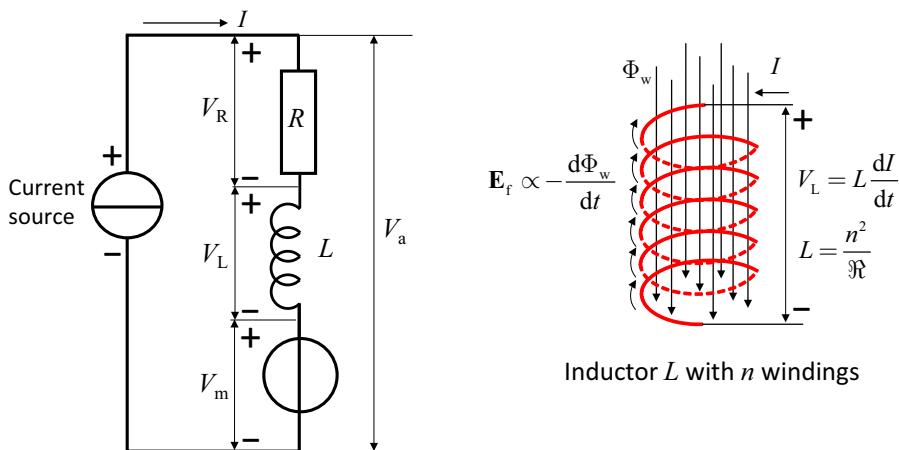
1. Useful mechanical energy (work).
2. Stored magnetic energy.
3. Heat loss.

In Figure 5.27 the electrical equivalent circuit diagram of an electromagnetic actuator is shown, with a current source as the electric power input, a magnetic potential energy storage, represented by the *self-inductance*  $L$ , and a resistive element that represents the thermal power output and the motor voltage that is related to the velocity. The electric element that represents the self-inductance is called an *inductor* because it works on the induced magnetic field by a current.

The self-inductance is a very important dynamic property of an electromagnetic actuator. It has been given the unit Henry [H], named after the American physicist Joseph Henry, who approximately at the same time as Faraday, determined the relation between the induced voltage over a coil as function of a change of enclosed flux, following Faraday's third Maxwell equation.

The self-inductance is defined as the ratio between the total flux  $\Phi_{w,t}$  of all windings summed, and the current in a coil.

$$L = \frac{\Phi_{w,t}}{I} \quad [\text{H}] \quad (5.54)$$



**Figure 5.27:** Electrical equivalent circuit diagram of an electromagnetic actuator.

The resistor  $R$  represents the resistivity of the windings and induced eddy-current losses, the inductor  $L$  represents the stored magnetic energy and the voltage source  $V_m$  is the induced voltage by the velocity. The right drawing shows the working principle of the inductor. It transforms a current change into an electromotive force that opposes to the external voltage, that causes the current change.

The total flux  $\Phi_{w,t}$  is equal to  $n\Phi_w$ , because a coil consists of  $n$  windings, and each winding encloses the same flux  $\Phi_w$ . The logic behind this relation of the self-inductance is based on Faraday's law, stating that the induced Electric field  $\mathbf{E}_f$  in a winding is equivalent to a change in flux over the closed surface inside the winding. This creates a proportional electromotive force over the winding. When each of the windings in a coil gets the same change of flux, the total coil would show an induced electromotive forced, that is the sum of the electromotive force of each individual winding as they are placed in series. Due to the minus sign in Faraday's law, the electromotive force within the coil is directed from the negative to the positive electrode, as explained in Chapter 2. This direction is opposite to the current direction so, when the electrodes are defined as in Figure 5.27, the induced voltage  $V_L$  at the electrodes of the inductor is equal to this electromotive force:

$$V_L = \mathcal{F}_e = n \frac{d\Phi_w}{dt} = \frac{d\Phi_{w,t}}{dt} \quad (5.55)$$

With the defined relation of  $L$ , the total current induced flux equals  $\Phi_{w,t} = LI$ , resulting in:

$$V_L = L \frac{dI}{dt} \quad (5.56)$$

It is to be expected that the self-inductance is determined by the dimensions and magnetic properties of the coil. In terms of number of windings and reluctance, this relation can be derived with Hopkinson's law of magnetics and is written as follows:

$$L = \frac{\Phi_{w,t}}{I} = n \frac{\Phi_w}{I} = n \frac{nI}{I\mathfrak{R}} = \frac{n^2}{\mathfrak{R}} \quad (5.57)$$

The implicit consequences of the self-inductance of a coil in the dynamic domain will be presented further in Section 5.4.1. In the present section it will be used to calculate the reluctance force in a five step approach based on the law of conservation of energy.

**Step 1: Power input** Using the equivalent circuit diagram of Figure 5.27, at any moment in time the electrical power is equal to:

$$P_{in}(t) = V_a(t)I(t) \quad (5.58)$$

For simplification the time dependency term ( $t$ ) is omitted in the further derived equations.

The total voltage  $V_a$  consists of three parts determined by the different elements of the equivalent circuit diagram:

$$V_a = V_R + V_L + V_m \quad (5.59)$$

The voltage over the resistor equals:

$$V_R = IR \quad (5.60)$$

The induced voltage over the self-inductance due to the change of the current equals:

$$V_L = L \frac{dI}{dt} \quad (5.61)$$

The motion voltage is the induced voltage due to the change of the flux by the movement only. In fact the change of the flux  $\Phi_w$  consists of two parts. One part is related to a change in the current in the coil and matches with the self-inductance. The other part is related to a change in the position:

$$V_m = n \frac{\partial \Phi_w}{\partial x} \frac{dx}{dt} \quad (5.62)$$

so that:

$$P_{in} = I^2 R + IL \frac{dI}{dt} + nI \frac{\partial \Phi_w}{\partial x} \frac{dx}{dt} \quad (5.63)$$

**Step 2: Power storage** Energy can be stored in only one element, the self-inductance. The stored magnetic energy is equal to the electric energy needed to create it, so it is calculated by integrating the power that was necessary to insert a current in the self-inductance. This power is equal to the current times the voltage caused by the change of current:

$$P_L = I(t)V_L = I(t)L \frac{dI}{dt} \quad (5.64)$$

If the current at  $t_0$  equals 0 and at  $t_1$  equals  $I_1$  the stored magnetic energy  $E_L$  at current level  $I_1$  for a certain value of  $L$  is calculated by integrating the power over the time from  $t_0$  to  $t_1$ :

$$E_L = \int_{t_0}^{t_1} P_L(t) dt = \int_{t_0}^{t_1} I(t)L \frac{dI}{dt} dt = L \int_0^{I_1} I(t) dI = \frac{1}{2} L I_1^2 \quad (5.65)$$

In case the self-inductance is depending on the position  $x(t)$  and the level of  $I_1$  is again the variable  $I$ , this can be written more generic as:

$$E_L = \frac{1}{2} L(x(t)) I^2 \quad (5.66)$$

The total power flowing into this storage element at any moment is the time derivative of this energy:

$$\begin{aligned} P_{\text{storage}} &= \frac{d}{dt}(E_L) \\ &= IL \frac{dI}{dt} + \frac{1}{2} I^2 \frac{dL(x(t))}{dt} \end{aligned} \quad (5.67)$$

**Step 3: Power output** Power can exit the system in only two ways: useful as mechanical power or useless as dissipated heat.

$$P_{\text{out}} = P_{\text{diss}} + P_{\text{mech}} \quad (5.68)$$

The dissipated heat equals:

$$P_{\text{diss}} = I^2 R \quad (5.69)$$

The mechanical power is determined in the next step.

**Step 4: Power balance** The total power balance becomes:

$$\begin{aligned} P_{\text{storage}} &= P_{\text{in}} - P_{\text{out}} \\ \left\{ IL \frac{dI}{dt} + \frac{1}{2} I^2 \frac{dL(x(t))}{dt} \right\} &= \left\{ I^2 R + IL \frac{dI}{dt} + nI \frac{\partial \Phi_w}{\partial x} \frac{dx}{dt} \right\} - \{I^2 R + P_{\text{mech}}\} \end{aligned} \quad (5.70)$$

After some elimination of terms, the mechanical output power becomes:

$$P_{\text{mech}} = nI \frac{\partial \Phi_w}{\partial x} \frac{dx}{dt} - \frac{1}{2} I^2 \frac{dL(x(t))}{dt} \quad (5.71)$$

**Step 5: Force** Mechanical power is delivered when a force is exerted during motion:

$$P_{\text{mech}} = Fv = F \frac{dx}{dt} \quad (5.72)$$

or

$$\begin{aligned} F &= P_{\text{mech}} \frac{dt}{dx} \\ &= nI \frac{\partial \Phi_w}{\partial x} - \frac{1}{2} I^2 \frac{dL(x)}{dx} \end{aligned} \quad (5.73)$$

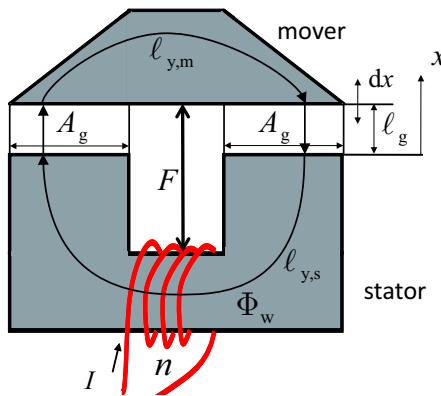
In this equation the first term is the linear relation of the force to the current, that is always present in any actuator. The second term is the squared relation of the force to the current and is caused by the magnetic energy that is stored in the self-inductance. In the Lorentz actuator of Section 5.3.1.2 the self-inductance is independent of the position because there is no surrounding ferromagnetic material that can influence the reluctance of the magnetic field from the coil. In that configuration the second term is zero and the general equation reduces to the previously derived Equation (5.43).

$$F = nI \frac{d\Phi_w}{dx} \quad (5.74)$$

In the following it will be shown that for a reluctance actuator without any permanent magnets the situation is quite different, because in that case the second term becomes exactly half of the first term which means that half the useful electrical energy (excluding resistive losses) is used to “charge” the coil with magnetic energy and only the other half is used for mechanical energy.

### 5.3.3 Variable reluctance actuator.

An example of an actuator working only on the reluctance force is the variable reluctance actuator of Figure 5.28. It is also just called a “reluctance actuator”. The principle is most widely applied as a basic electromagnet, that is created with a current-carrying coil around a ferromagnetic yoke. This electromagnet attracts other pieces of ferromagnetic material and most



**Figure 5.28:** The variable reluctance actuator only works by the principle of the reluctance force. The moving part is attracted with a force proportional to the current squared and inversely proportional to the distance squared.

people know it from the magnet that pulls-up a car with a helicopter in a crime movie. To calculate the force, Equation (5.73) is used, where the relative movement of the mover to the stator equals  $dx$ :

$$F = nI \frac{\partial \Phi_w}{\partial x} - \frac{1}{2} I^2 \frac{dL(x)}{dx} \quad (5.75)$$

In this case, without a permanent magnet,  $\partial \Phi_w / \partial x$  is only determined by the current and directly related to the self-inductance, that in its turn is only a function of the position:

$$\frac{\partial \Phi_w}{\partial x} = \frac{d}{dx} \frac{LI}{n} \quad (5.76)$$

Now these two equations combine to:

$$F = I^2 \frac{dL(x)}{dx} - \frac{1}{2} I^2 \frac{dL(x)}{dx} = \frac{1}{2} I^2 \frac{dL(x)}{dx} \quad (5.77)$$

The self-inductance is determined by the series reluctance of the two ferromagnetic parts, the *stator* and the *mover*, with a total length  $\ell_y = \ell_{y,s} + \ell_{y,m}$  and the two air-gaps with a total length of  $2\ell_g$ , because the magnetic flux has to cross the air-gap twice.

$$L = \frac{\Phi_t}{I} = \frac{n^2}{\mathfrak{R}_c + \mathfrak{R}_g} = \frac{n^2}{\frac{\ell_y}{A_y \mu_0 \mu_r} + \frac{2\ell_g}{A_g \mu_0}} \quad (5.78)$$

With  $\mu_r \gg 1$  and relative large values of  $\ell_g$ , this can be approximated<sup>2</sup> into:

$$L \approx n^2 \frac{A_g \mu_0}{2\ell_g} \quad (5.79)$$

With this value for the self-inductance, Equation (5.77) with  $x = \ell_g$  gives after a bit of algebra:

$$F \approx - \left( \frac{nI}{\ell_g} \right)^2 \frac{\mu_0 A_g}{4} \quad (5.80)$$

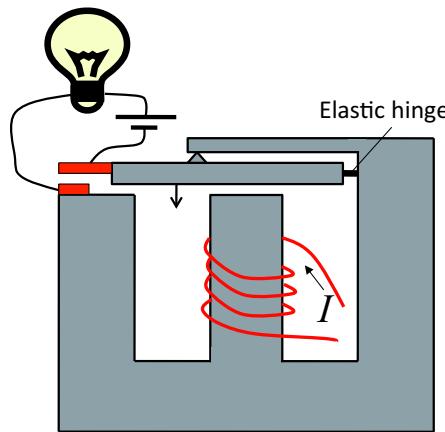
The minus sign indicates that the force on the mover is directed in the negative  $x$  direction. This corresponds with the known “pulling” direction of an electromagnet. It is also clear from this equation that the reluctance actuator is extremely non-linear. First of all the force only works in one direction and the “gain”, that equals the force to current ratio, increases with higher currents. But also the dependence on a movement in the  $x$  direction is high and non-linear which implies a significant stiffness. The magnitude of the force decreases when the displacement increases. This means, that the actuator has a negative non-linear stiffness. When saturation effects are neglected, an infinitely small gap could create an infinite force. In reality the flux density is limited to the saturation of the applied ferromagnetic material, which corresponds for ferromagnetic (soft) iron to a value in the order of 2 T. The resulting force can then be calculated with the equations of the next section.

### 5.3.3.1 Electromagnetic relay

An electromagnetic relay, as schematically shown in Figure 5.29, is a well-known application of a reluctance actuator. The squared force relation of the reluctance force is optimally utilised in this electrically activated switch. A relay consists of a stator from a ferromagnetic material provided with a coil and a mover which is connected via an elastic hinge to the stator. The elastic hinge is pre-stressed so that the mover is pushed to its limit with a certain force. An electrical contact is made when the mover is attracted to the stator. When a current starts flowing in the coil, at a certain current level the reluctance force equals the pre-stress force and the mover will start moving towards the stator. This movement decreases the air-gap so the force will increase further. This results in a non-linear exponential avalanche effect that creates a very strong force and a fast closing of the contacts.

---

<sup>2</sup>Later it will be shown that this approximation can cause large errors with high levels of flux density and small values of  $\ell_g$ !



**Figure 5.29:** An electromagnetic relay is a bistable system. the switch is pulled towards the end stop by a pre-stressed elastic hinge. When the current surpasses a certain threshold the magnetic force is larger than the pre-stress and the switch will move. This movement reduces the reluctance of the magnetic field with a resulting increase in force. The “avalanche” effect will close the switch at a high speed.

### 5.3.3.2 Force exerted by a magnetic field

With the force Equation (5.80) of a reluctance actuator also the magnetic flux density  $B_g$  in an air-gap by a permanent magnet, or any other source, can be related to the force that is exerted between the two sides of the air-gap. In the reluctance actuator, the relation between the magnetic flux density in the air-gap and the current is derived from the total flux that passes the air-gap. It has a clear relation with the self-inductance  $L$ :

$$B_g = \frac{\Phi_w}{A_g} = \frac{LI}{nA_g} \quad (5.81)$$

The self-inductance value  $L$  of the reluctance actuator was obtained with Equation (5.79):

$$L \approx n^2 \frac{A_g \mu_0}{2\ell_g} \quad (5.82)$$

This gives with Equation (5.81):

$$B_g \approx \frac{nI\mu_0}{2\ell_g} \Rightarrow nI \approx \frac{2B_g\ell_g}{\mu_0} \quad (5.83)$$

This can be combined with Equation (5.80):

$$F \approx -\left(\frac{nI}{\ell_g}\right)^2 \frac{\mu_0 A_g}{4} \quad (5.84)$$

giving with a bit algebra:

$$F \approx -\frac{B_g^2 A_g}{\mu_0} \quad (5.85)$$

The total gap surface is  $2A_g$  so a magnetic flux density  $B$  causes an equivalent virtual pulling “pressure”  $P_m$  of:

$$P_m \approx \frac{F}{2A_g} \approx -\frac{B_g^2}{2\mu_0} \quad (5.86)$$

This means for example that a magnetic flux density of 1 T creates a pulling force per unit of surface (negative pressure) of:

$$P_{1T} \approx \frac{1}{8\pi \cdot 10^{-7}} \approx 0.4 \text{ [MPa]} \quad (5.87)$$

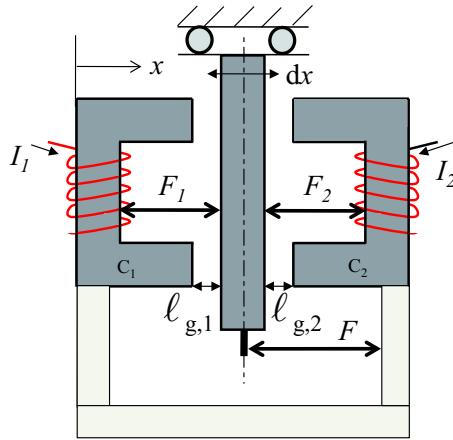
This equals a negative pressure of 4 times vacuum, which immediately give the reason why permanent magnets are often used for a contact-less pre-load of air-bearings as applied in precision positioning systems.

### 5.3.4 Hybrid actuator

It is valuable to investigate how a reluctance actuator can be made suitable for controlled positioning systems, because of the high force-to-current ratio in configurations where a very small air-gap is allowed. The first problem to solve is the unidirectional force.

#### 5.3.4.1 Double variable reluctance actuator

The limitation by the unidirectional force can either be solved by providing a passive force in the other direction by means of a spring or by combining two reluctance actuators into one actuator, like shown in Figure 5.30. This configuration proves to be a very interesting system as will follow from the calculations. First of all the force on the central mover equals the difference of the forces of each actuator. In the previous example of the single reluctance actuator the positive  $x$  direction was defined from the stator upwards to the mover resulting in a minus sign for the force. In this example the positive  $x$  direction is defined in the right direction as shown in the picture. This means that the force that acts on the mover by the left actuator half ( $F_1$ ) has a negative sign and the force by the right actuator half ( $F_2$ ) has a positive sign. Using Equation (5.80) twice, for each half of the actuator, the total



**Figure 5.30:** A double reluctance actuator with a central mover shows a linear current to force relationship in the middle position. This is caused by the balancing of the force of the two non-linear actuators.

force becomes:

$$F = F_2 - F_1 = \left( \frac{nI_2}{\ell_{g,2}} \right)^2 \frac{\mu_0 A_g}{4} - \left( \frac{nI_1}{\ell_{g,1}} \right)^2 \frac{\mu_0 A_g}{4} = \frac{n^2 \mu_0 A_g}{4} \left( \frac{I_2^2}{\ell_{g,2}^2} - \frac{I_1^2}{\ell_{g,1}^2} \right) \quad (5.88)$$

This shows as a first conclusion, that in the mid position, where  $\ell_{g,1} = \ell_{g,2}$ , the force is zero when the currents are equal. As a next step it is interesting to see what happens if the currents in both halves are modulated, while keeping the sum of the currents constant. This means that current  $I_1$  becomes  $I_a - \Delta I$  and current  $I_2$  becomes  $I_a + \Delta I$ , while  $I_a$  is the average current.

The calculations start with Equation (5.88), with the mover in the mid position,  $\ell_{g,1} = \ell_{g,2} = \ell_g$ .

$$F = \frac{n^2 \mu_0 A_g}{4} \left( \frac{I_2^2 - I_1^2}{\ell_g^2} \right) \quad (5.89)$$

With  $I_1 = I_a - \Delta I$  and  $I_2 = I_a + \Delta I$  the force becomes:

$$F = \frac{n^2 \mu_0 A_g}{4} \left( \frac{(I_a + \Delta I)^2 - (I_a - \Delta I)^2}{\ell_g^2} \right) = \frac{n^2 \mu_0 A_g}{4} \left( \frac{4I_a \Delta I}{\ell_g^2} \right) \quad (5.90)$$

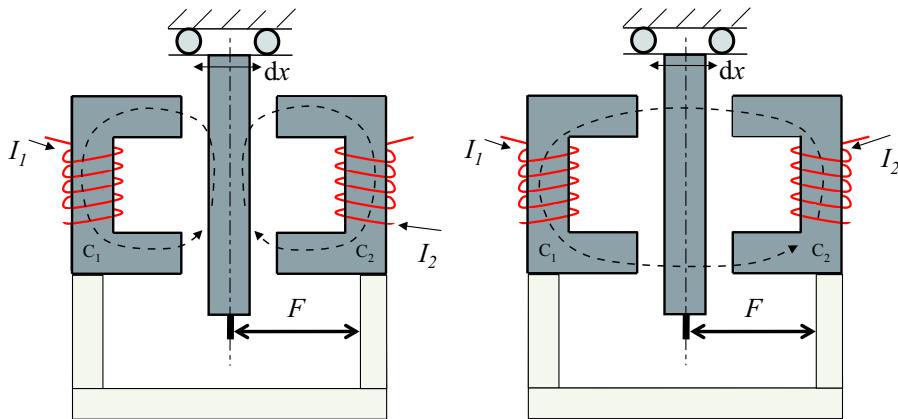
With this result, the force to current ratio of this actuator for the current change  $\Delta I$  can be written as:

$$\frac{F}{\Delta I} = \frac{I_a n^2 A_g \mu_0}{\ell_g^2} \quad (5.91)$$

This means that the force to current ratio is depending on the level of the current level that is common for both halves. It gives the possibility to control the differential gain by this common current level! The second conclusion is that the force is linearly related to the differential current at the same mid-position! This means that this actuator can be used in a linear control system with a separately controllable gain. Unfortunately this is only the case at the mid position. At other positions the actuator half with the smallest gap will dominate the other half, and the linearity gradually disappears. Still for small displacements with a high force this configuration can be useful.

Another drawback of this configuration is the *negative stiffness* that depends on the common current. This effect can be explained from the fact that at any position away from the mid-position the mover will mainly be attracted to the position where it is closest to the stator. This makes this actuator less suitable for situations, where a direct coupling between stator and mover is undesired like in a vibration isolation system. In principle this negative stiffness can be compensated with the positive stiffness of a mechanical spring, but in practice it is difficult to achieve a reliable reduction of more than a factor 10 due to tolerances.

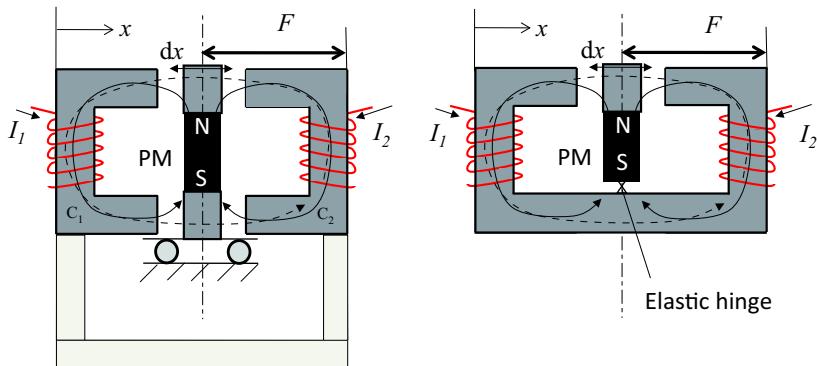
With the reluctance actuator it was shown that by combining two reluctance actuators the actuator can be linearised by providing it with a current that is common for both halves. This principle has as drawback that a DC current will contribute to the power losses in the actuator and also the negative stiffness is quite large. It is a logical step to replace the flux created by this common mode current by the flux of a permanent magnet and thus combine the best of both worlds: the high force of the reluctance actuator with the linearity of the permanent magnet actuator. To introduce a permanent magnet in the double reluctance actuator it is necessary to carefully consider the working principle as explained previously. First of all the force is generated by the flux in the air-gap. Because of the squared relation between flux and force, a force is obtained that is equal to the multiplication of the flux related to the common current and the flux related to the differential current. This means that a perfectly linear very strong actuator is obtained under the following two conditions. First it must be possible to create a common flux in the air-gap by means of a permanent magnet that is independent of the position of the mover. Secondly the permanent magnet is not allowed to introduce a significant additional reluctance for the flux from the current in the coil. In the following section it is shown how this can be achieved.



**Figure 5.31:** The flux in the double reluctance actuator can be changed by inverting the current direction in one of the coils. Because of the squared relation in the reluctance force this inversion has no influence on the forces. Nevertheless the flux path is completely changed. This creates the possibility to change the material properties of the mover without impacting the reluctance force.

#### 5.3.4.2 Combining two sources of magnetic flux

The first step is to be aware that the double reluctance actuator of Figure 5.30 works independent of the current direction in the coils due to the squared relation between current and force. Though it has no impact on the force, the flux is different if the current in one coil is reversed as shown in Figure 5.31. If the flux in both coils is in the same direction (upwards or downwards) the flux of both coils has to flow back vertically through the mover. When the current in one of the coils is reversed, as shown in the right picture, the flux of both coils share the same path which only crosses the gap. The vertical path in the mover is avoided which means that the middle part of the mover can be replaced by another non-ferromagnetic element like a permanent magnet. This is shown in Figure 5.32 where the flux of the permanent magnet perfectly combines with the induced flux of the current without adding additional reluctance to the latter. In the right picture a simplified version is shown where one side of the mover is connected to the stator by means of an elastic hinge that enables rotation in combination with a low reluctance path for the permanent magnet flux. This elastic hinge can be used to partly compensate the still not completely avoidable inherent negative stiffness of the actuator as will be explained later. It is clear that the permanent magnet creates an unstable situation because



**Figure 5.32:** By inserting a permanent magnet in the mover of the double reluctance actuator, an additional flux is created, that can replace the flux that was generated by the average current of the original double reluctance actuator. This replacement strongly reduces the power consumption of the actuator.

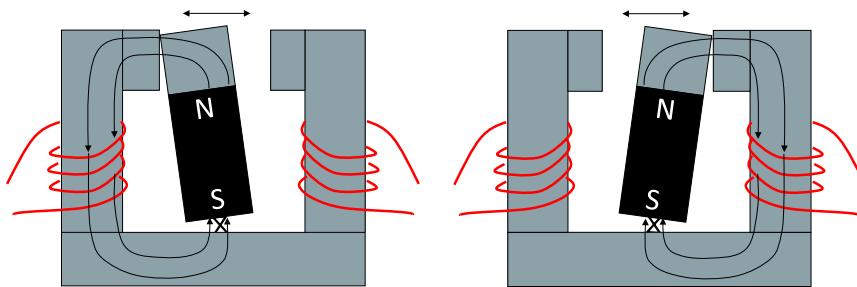
the magnetic circuit has its lowest reluctance at one of the two positions as shown schematically in Figure 5.33. With the mover in the middle position the flux is equally divided over the two half's of the actuator and the attraction force in both air-gaps cancel each other out. As soon as the mover moves in one of the two directions the flux will increase in that half and decrease in the other which will cause the negative stiffness. Figure 5.34 is used to calculate the flux of the permanent magnet. While the reluctance for each gap is proportional to their length  $x$ , the total reluctance of both gaps combined is calculated by taking the reluctance of both gaps *in parallel* according to the following relation.

$$\frac{1}{\mathfrak{R}} = \frac{1}{\mathfrak{R}_1} + \frac{1}{\mathfrak{R}_2} \quad (5.92)$$

With  $dx$  the displacement of the mover,  $\ell_{g,1} = \ell_g + dx$  and  $\ell_{g,2} = \ell_g - dx$  being the length of both gaps,  $A_g$  the cross section of the gaps perpendicular to the flux and the ferromagnetic material having an infinite  $\mu_r$ , the following equation is obtained:

$$\mathfrak{R} = \frac{1}{\mu_0 A_g \left( \frac{1}{\ell_g + dx} + \frac{1}{\ell_g - dx} \right)} = \frac{\ell_g^2 - (dx)^2}{2\mu_0 A_g \ell_g} \quad (5.93)$$

Around the mid position with  $dx \ll \ell_g$ , this reluctance is approximately constant but at the outer positions this causes the total reluctance to decrease. This leads to a higher flux and stronger negative stiffness, however to a



**Figure 5.33:** The permanent magnet flux division in a hybrid actuator depends on the position of the mover.

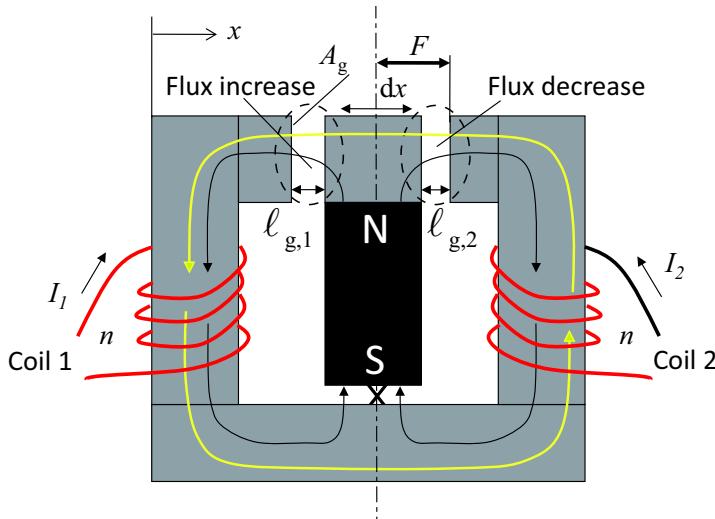
lesser extent than with the double reluctance actuator from the previous section. This is caused by the fact that the reluctance of the permanent magnet field is dominated by the permanent magnet material itself which limits the flux density to below  $B_r$ . Remember that a modern permanent magnet acts like an air coil with a very large current. With the reluctance actuator a small gap will lead to a steep increase of flux density that would only be limited by the saturation. Because of this lower and more linear negative stiffness of the hybrid actuator it can more easily be compensated by a positive spring, which in this configuration is determined by the elastic hinge. In Figure 5.34 the effect of a current through the coils is shown. The flux from each coil will follow the path of lowest reluctance, according to the yellow arrows, showing that it will not pass the permanent magnet but go round in the ferromagnetic part. As mentioned the winding direction of the coils is such that their fluxes positively add together.

Depending on the current direction, the combination of the flux of the coils with the flux of the permanent magnets results in a lower flux density in one air-gap and a higher flux density in the other air-gap, causing an attracting force in the direction of the increased flux density. As this happens in two directions, the force is depending on the current direction and the effect is linearised according to the same reasoning as with Equation (5.91) for the double reluctance actuator.

### 5.3.4.3 Hybrid force calculation

In order to approximate the magnitude of the force, first Equation (5.73), the general equation of electromagnetic actuators is applied.

$$F = nI \frac{d\Phi_w}{dx} - \frac{1}{2} I^2 \frac{dL(x)}{dx} \quad (5.94)$$



**Figure 5.34:** The flux in a hybrid actuator is a combination of the flux of the permanent magnet and the flux of the coils. In one air-gap the flux of the coils has the same direction as the flux of the permanent magnet, increasing the total magnetic field, while at the other air-gap it is just opposite. This results in a strong net force in the direction of the strongest magnetic field.

The self-inductance of the coils is hardly influenced by the position of the mover because the reluctance of the coils consists of the reluctance of the two air-gaps *in series*.

$$\mathfrak{R} = \mathfrak{R}_1 + \mathfrak{R}_2 \quad (5.95)$$

When one gap gets smaller the other gets bigger so the second term can be neglected and the known relation from the Lorentz actuator is used:

$$F = nI \frac{d\Phi_w}{dx} \quad (5.96)$$

In this equation it is necessary to replace  $n$  by  $2n$  because the flux of the windings of both coils are added together as they are working in the same direction. The flux through the windings  $\Phi_w$  consists of the flux of the permanent magnet and the flux caused by the current through the coils. Under condition of a non-saturated iron part, as mentioned in Equation (5.95), the flux caused by the current is determined by the reluctance of the air-gaps in series. The total reluctance of these gaps is equal to:

$$\mathfrak{R} = \frac{\ell_{g,1} + \ell_{g,2}}{\mu_0 A_g} \quad (5.97)$$

This is a constant because the sum of both gap lengths is constant. This means  $d\Phi_w/dx$  is only determined by the permanent magnet flux. The force can be approximated by taking the two extreme positions when the mover is just hitting the stator. In both situations the iron determines a low reluctance path for the permanent magnet, which means that the flux density  $B_m$  of the permanent magnet will approach  $B_r$ . So by changing the position from one side to the other the average flux changes approximately from zero to  $B_r A_m$ . This gives:

$$\frac{d\Phi_w}{dx} = \frac{B_m A_m}{\ell_{g,1} + \ell_{g,2}} \approx \frac{B_r A_m}{\ell_{g,1} + \ell_{g,2}} \quad (5.98)$$

With the two equal coils ( $2nI$ ) this results in an estimated force to current<sup>3</sup> ratio of:

$$\frac{F}{I} \approx 2n \frac{d\Phi_w}{dx} \approx 2n \frac{B_r A_m}{\ell_{g,1} + \ell_{g,2}} \quad (5.99)$$

Like all actuators, also the hybrid actuator has a position dependency of the force, not only related to the mentioned negative stiffness, but also due to the position dependent force to current ratio. In the hybrid actuator this second effect is caused by the changing total reluctance of the permanent magnet circuit as function of the position. As discussed earlier, the reluctance for the permanent magnet at the mid position equals the the reluctance of the two air-gaps in parallel. In the outer positions one of the air-gaps is very small, resulting in a lower total reluctance. This means that the flux of the permanent magnet increases at the outer positions, giving a higher value of  $d\Phi_w/dx$  than in the mid position.

It is clear that these calculations are very approximative due to the complex nature of the system, with potential saturation of the iron part and the approximation of the permanent magnet flux in the extreme positions.

To check this approximation, the force in the mid position can be calculated in another way, by means of Equation (5.86) with the cross section of the air-gaps and an estimation of the flux of the permanent magnet.

Like with the double reluctance actuator two counteracting forces are combined. The force in air-gap 1 works in the negative  $x$  direction and the force in air-gap 2 in the positive  $x$  direction. The flux density in the air-gaps ( $B_g$ ) equals the combination of the flux density by the permanent magnet ( $B_{g,m}$ ) and the flux density caused by the coil ( $B_{g,c}$ ). With the current direction as defined in Figure 5.34 the total flux density in air-gap 1 equals  $B_{g,m} + B_{g,c}$

---

<sup>3</sup>Here the current ( $I$ ) represents the variable current to control the force in a position controller. It is comparable with the  $\Delta I$  from Equation (5.91)

and in air-gap 2 it equals  $B_{g,m} - B_{g,c}$  and the following equation is obtained:

$$F = F_2 - F_1 = \frac{A_g}{2\mu_0} (B_{g,m} - B_{g,c})^2 - \frac{A_g}{2\mu_0} (B_{g,m} + B_{g,c})^2 = -\frac{2A_g B_{g,m} B_{g,c}}{\mu_0} \quad (5.100)$$

The minus term is due to the fact that a positive current in the shown direction results in a force on the mover to the left which is the negative  $x$  direction.

In case the air-gap cross section is large with respect to the length of the air-gap, the leakage flux will be moderate ( $\lambda > 0.80$ ). With a well designed system, the flux density of the magnet is chosen to be  $B_r/2$  and the flux is divided over two halves which means that a flux density in each gap is equal to:

$$B_{g,m} = \frac{0.8 \cdot B_r}{2} \frac{A_m}{2A_g} \quad (5.101)$$

When the reluctance of the air-gaps is large in comparison with the reluctance of the ferromagnetic parts, the flux induced by the current equals (two coils with  $n$  windings!):

$$B_{g,c} = \frac{\Phi_w}{A_g} = \frac{2nI}{A_g R} = \frac{2nI\mu_0}{\ell_{g,1} + \ell_{g,2}} \quad (5.102)$$

Combining these equations for  $B_{g,m}$  and  $B_{g,c}$  with Equation (5.100) gives the following force to current ratio of the hybrid actuator:

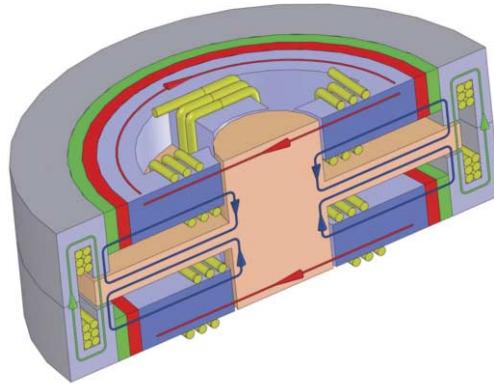
$$\frac{F}{I} = 0.8n \frac{B_r A_m}{\ell_{g,1} + \ell_{g,2}} \quad (5.103)$$

When this result is compared with Equation (5.99) it clearly shows more than a factor two decrease of the force in the mid position which is caused by the reluctance of the air-gaps for the permanent magnet flux and the choice to use as little as possible magnet material ( $B_m = B_r/2$ ). In practical designs often a larger magnet is chosen which brings the force to a higher value and reduces the position dependency of the force to current ratio because the flux density will be more constant closer to  $B_r$ .

It can also be concluded that with increasing complexity of the magnetic circuit the need for more exact calculations using FEM computer simulations is more strong. Nevertheless the presented calculations provide a good sense for the order of magnitude the real forces in a practical actuator.

#### 5.3.4.4 Magnetic bearings

A nice application example of a hybrid actuator is its use in a magnetic bearing. Figure 5.35 shows a fully integrated magnetic bearing that supports



**Figure 5.35:** Five degrees of freedom homo-polar magnetic bearing. The flux by the current in the coils (red arrows) is guided via a different route than the flux from the permanent magnets (blue arrows). In the air-gaps they combine to higher or lower values depending on the current direction, thereby enabling a change in the force in different directions.

a fast-rotating shaft in five directions, three orthogonal linear directions and two rotations. Only the rotation around the central axis is left free. This concept is designed for a high speed micro-milling centre that has to operate at speeds above 300.000 rpm. Also in this system the permanent magnets create a bias flux which is modulated by the coils. By tracing the flux of the different coil sections as indicated with the red arrows and combine these with the flux from the permanent magnet as indicated with the blue arrows it can be imagined how all forces are created. It is worthwhile to notice, that also in this case the permanent magnets do not increase the reluctance of the flux from the coils. Still the configuration is essentially different from the example in the previous section as it uses the low reluctance path in the third dimension for the flux induced by the current in the coils. This is just another illustration of the large amount of configurations, that are possible in electromagnetic actuators. This example of a magnetic bearing is called *homo-polar* because the permanent magnet flux flows in the axial direction through the rotating shaft, which means that the flux in the shaft does not change due to the rotation. As a consequence of this configuration, power consuming eddy-currents are prevented that would otherwise be induced by the changing flux in the rotating shaft. An important property of a magnetic bearing is its inherent negative stiffness. The shaft needs to rotate freely which means that this stiffness can not be compensated by a mechanical spring like the elastic hinge of the previously presented linear

hybrid actuator. In Chapter 4 it is shown how negative feedback control enables to stabilise the position of the shaft inside the magnetic bearing by providing a virtual spring with positive stiffness and damping. This active control of magnetic bearings has enabled their use in critical applications that do not allow mechanical contact and require strict control of deviations from the ideal position of the shaft at often extremely high rotation speeds.

## 5.4 Application of electromagnetic actuators

The application of the three presented electromagnetic actuators in mechatronic systems is determined by their properties in relation to the other parts of the system. This section will first present the electrical interface with the amplifier followed by a comparison of the three types with some realistic data to illustrate their different characteristics.

### 5.4.1 Electrical interface properties

In mechatronic positioning systems the actuator is always used in combination with an amplifier. This amplifier has also its limitations as will be presented in Chapter 6. The impedance of the actuator strongly influences the total transfer function and stability of the amplifier-actuator combination and also the amplifier output characteristics strongly influences the behaviour of the actuator. In this section these effects will be elaborated. First the dynamic effect of the self-inductance will be investigated in the frequency domain, when the amplifier can be approximated as an ideal current source. This means that the behaviour of the actuator does not influence the current of the amplifier. It will be shown that these dynamic properties will only depend on the dimensions of the coil and not on the number of windings and the wire thickness.

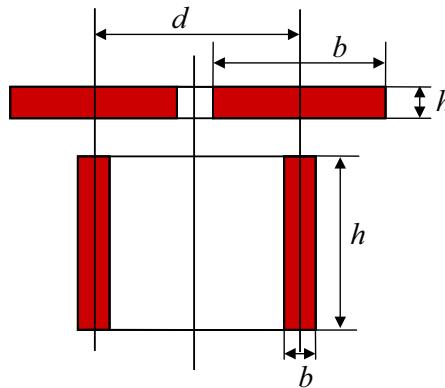
Directly related to the self-inductance is the ability of the actuator to change the acceleration rapidly. This is represented by the *jerk*., that is the derivative of acceleration over time.

As a last step the amplifier will be modelled in a more realistic way, where the behaviour of the actuator influences the current. It will become clear, that the actuator shows damping properties, when the amplifier is not an ideal current source.

#### 5.4.1.1 Dynamic effects of self-inductance

In Figure 5.27 it was shown that an electromagnetic actuator can be represented in the electrical domain as a series of a resistance, a self-inductance and a voltage source. When the actuator is supplied by a current, the voltage over the actuator will be equal to:

$$V_a = V_m + IR + L \frac{dI}{dt} \quad (5.104)$$



**Figure 5.36:** Two different coil configurations with their basic dimensions shown as an example of the wide amount of possibilities. Next to these rotation symmetrical shapes, also square, race-track and even three dimensional shapes are used.

In order to get an idea of real values, the resistance from the windings is calculated with the following relation:

$$R = \rho_r \frac{n \ell_w}{A_w} \quad (5.105)$$

with

$\rho_r$  = resistivity of the wiring material. [Ω/m]

$\ell_w$  = average length per winding. [m]

$n$  = number of windings.

$A_w$  = cross section of the wire. [m<sup>2</sup>]

When designing a coil of a certain size, it mostly starts with the dimensions, giving the average length of each winding ( $\ell_w = \pi d$ ) and its cross section ( $A_{c,w} = bh$ ) as shown for two typical coil configurations in Figure 5.36.

$A_{c,w}$  relates to  $A_w$  in the following manner:

$$A_w = \frac{\gamma A_{c,w}}{n} \quad (5.106)$$

With  $\gamma$  being the fill factor which reduces the useful volume of the coil due to round windings and the insulation. Practical values for  $\gamma$  range between approximately 0.5 for thin round wires to 0.9 for flat wires. With these factors the resistance becomes:

$$R = \frac{\rho_r n^2 \ell_w}{\gamma A_{c,w}} \quad (5.107)$$

Assuming  $\gamma$  and  $\rho_r$  are constant for a given coil size, the resistance is proportional to the number of windings squared.

To estimate the effect of the self-inductance, Equation (5.104) is used in the mechanical stationary situation, which means that  $V_m = 0$

$$V_a(t) = IR + L \frac{dI}{dt} \quad (5.108)$$

After applying the Laplace transform to bring the equation to the frequency domain the following relation is found:

$$V_a(s) = I(R + sL) \implies V_a(\omega) = IR \left(1 + \frac{j\omega L}{R}\right) \quad (5.109)$$

To give a value to the ability to change the force rapidly, the electrical time constant  $\tau_e$  is introduced, defined as:

$$\tau_e = \frac{1}{\omega_0} = \frac{L}{R} \quad (5.110)$$

With Equation (5.109) the impedance  $Z$  becomes:

$$Z(\omega) = \frac{V_a}{I} = R(1 + j\omega\tau_e) \quad (5.111)$$

This makes clear, that above  $\omega_0 = 1/\tau_e$ , the impedance, and consequently also the voltage that is necessary to drive a current to the actuator, increases proportional with a slope of +1 as function of the frequency. This entails a heavy requirement for the amplifier that has to deliver both a high voltage to cope with the self-inductance and a high current to deliver the force at elevated frequencies.

For this reason it is important to keep  $\tau_e$  as low as possible. When the resistance  $R$  is replaced by Equation (5.107) and Equation (5.57) is used to replace the self-inductance  $L$  by  $n^2/\mathfrak{R}$ , the following relation for the electrical time constant is obtained:

$$\tau_e = \frac{\gamma A_{c,w}}{\mathfrak{R} \rho_r \ell_w} \quad (5.112)$$

This shows that  $\tau_e$  can only be tuned by the dimensions of the coil, the resistivity and the reluctance.

#### 5.4.1.2 Limitation of the “jerk”

In high speed precision controlled positioning systems, the voltage that is required to change the current in the self-inductance of the actuator, needs

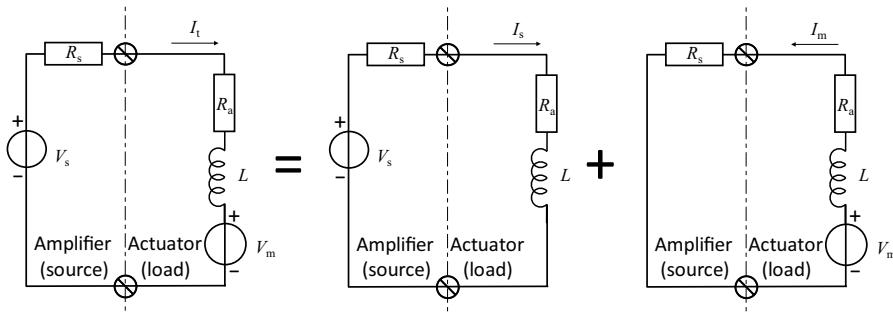
to be actively limited in order to stay within the maximum voltage range of the power amplifier. When exceeding this range, the amplifier does not reliably transmit the control information to the actuator. This in fact means that the system is (temporarily) out of control.

In most actuators the current is proportional to the force and as a consequence it is proportional to the acceleration of the driven mass. This means that by limiting the change of current, the change of acceleration, the “jerk” is limited. Because of this limitation, high speed precision controlled positioning systems need to control both position, speed, acceleration and jerk. Sometimes even the derivative of the jerk over time is controlled. This entity is called *snap*.

In case of fast moving reciprocating systems like wafer scanners it is clear that a limitation of the jerk automatically limits the maximum attainable acceleration as being the integral of the jerk over time. It is also for this reason that a Lorentz actuator is often the preferred choice in these systems as it possesses the lowest value of  $\tau_e$  from all actuator types, especially when the ferromagnetic part is completely omitted, like the example shown in Figure 5.26. It is however also true that a more efficient actuator based on the reluctance or hybrid principle would needs less current and this could compensate for the higher voltage that is necessary to realise the high  $dI/dt$  through the self-inductance. This is another area for optimisation in mechatronic systems.

#### 5.4.1.3 Damping caused by source impedance

In the previous modelling of the behaviour of an electromagnetic actuator it was always assumed that the current was given. In practice the current is supplied by a power amplifier. In the following chapter on electronics it will be explained that a power amplifier is not capable to deliver a current that is fully independent of the load. It will be shown that the output of the amplifier possesses a certain source impedance that influences the behaviour of the connected actuator, the load. As a result all electromagnetic actuators show a certain amount of damping, when connected to a power amplifier. This damping is caused by the motion voltage  $V_m$  that is induced in the coil, due to the change of flux that is related to the movement. This voltage in its turn will cause a current through the circuit that is further determined by the series of the impedance of the actuator and the source impedance of the amplifier as shown in Figure 5.37. In the following the effect of this current will be explained by means of the Lorentz actuator, because of its constant permanent magnet flux that simplifies the example.



**Figure 5.37:** Electrical circuit of an amplifier with a resistive output impedance  $R_s$ , that acts as a source for an electromagnetic actuator, represented by its electrical equivalent. The total current in the circuit  $I_t$  is the combination of the current  $I_s$ , caused by the amplifier and the current  $I_m$ , caused by the motion voltage  $V_m$ ,  $I_t = I_s - I_m$ . The damping is caused by  $I_m$  only.

In principle the effect of the two voltage sources can be combined linearly and the total current in the circuit is equal to the sum of the currents caused by both voltage sources.

$$I_t = I_s - I_m = \frac{V_s}{R_s + R_a + j\omega L} - \frac{V_m}{R_s + R_a + j\omega L} = \frac{V_s - V_m}{R_s + R_a + j\omega L} \quad (5.113)$$

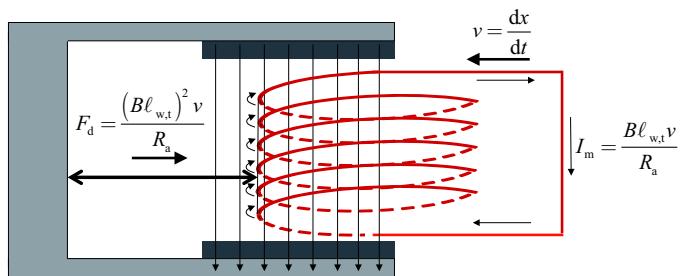
This means that it is allowed to analyse the behaviour of an actuator for each voltage source separately, while the other source is replaced by a conducting wire. The current caused by the amplifier voltage creates the force that was described in the previous sections. The current caused by the motion voltage is the new factor. It creates the damping effect by the resulting force in the actuator, counteracting the velocity. The related damping coefficient  $c$  can be derived as follows for a Lorentz actuator.

Using Figure 5.38 for the directions, it starts with the motion voltage:

$$V_m = (B\ell_{w,t})v \quad [\text{V}] \quad (5.114)$$

with  $\ell_{w,t}$  being the total length of the windings inside the magnetic field. To determine the resulting current, the impedance of the self-inductance is neglected in respect to  $R_a + R_s$ . This is mostly allowed with Lorentz actuators working at relatively low frequencies.

$$I_m = \frac{V_m}{R_s + R_a} = \frac{(B\ell_{w,t})v}{R_s + R_a} \quad [\text{A}] \quad (5.115)$$



**Figure 5.38:** A Lorentz actuator shows a velocity dependent damping force when the two terminals of the coil are connected by means of a low impedance circuit ( $R_s = 0$ ). Following Faraday's law, the induced electric field by the movement-related flux-change creates a current  $I_m$  in the direction as shown, due to the minus sign. This current generates a Lorentz Force  $F_d$  that is directed in the opposite direction of the movement according to the corkscrew rule.

As the current will flow in the same magnetic field, a Lorentz force will occur:

$$F_d = (B \ell_{w,t}) I_m \text{ [N]} \quad (5.116)$$

Filling in  $\frac{(B \ell_{w,t})v}{R_s + R_a}$  for  $I_m$  gives:

$$F_d = \frac{(B \ell_{w,t})^2 v}{R_s + R_a} \text{ [N]} \quad (5.117)$$

So the damping coefficient  $c$  is:

$$c = \frac{F_d}{v} = \frac{(B \ell_{w,t})^2}{R_s + R_a} \text{ [Ns/m]} \quad (5.118)$$

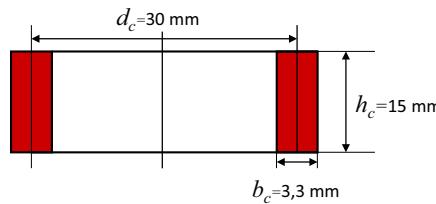
The conclusion of this equation is, that the damping is proportional to the  $B\ell$ -factor squared. This factor is also related to the force to current ratio. This means, that it is not preferred to reduce this  $B\ell$ -factor when less damping is needed. In that case it is better to control the damping by adapting the resistive value of the circuit.

In loudspeaker systems it is necessary to create damping, in order to suppress the resonance of the loudspeaker cone with the stiffness of its support and the air cabinet. For that reason amplifiers for music need to have a low output impedance. For precision mechatronic systems it is more often required to avoid the motion related damping as much as possible, because

of the transmissibility of external vibrations through the electromagnetic damper. For that reason in those cases the output impedance of the amplifier should be as high as possible. This means that the amplifier will act like a current source with a current level, that is almost independent of the load. As an additional effect at first sight the self-inductance will not cause any trouble in the dynamic performance of the system, but as will be shown in the next chapter the impedance of the self-inductance can interfere with the stability of the amplifier. It remains important to keep the electrical time constant  $\tau_e$  as low as possible.

### 5.4.2 Comparison of the actuation principles

In designing mechatronic systems, making choices is often rather difficult because of the many, sometimes contradictory, requirements. In practice, many decisions are made based on experience and personal preference, which can lead to the solutions like the mousetraps in the first chapter. This is often also the case with the choice of an electromagnetic actuator. Many people, especially those working with high velocity and extremely accurate positioning systems, almost without thinking choose a Lorentz actuator, because of its linearity and low values of mechanical stiffness, that could transfer vibrations from outside. In the application of the wafer scanners that are presented in Chapter 9, the inherent drawback of the limited force in relation to the electrical power has resulted in an immense power consumption up to several kilowatts per stage. In fact, the maximum acceleration of a Lorentz actuator is limited to a physical maximum that depend on the allowable thermal dissipation. Above a certain specific force per unit of mass, the necessary increase of mass to cope with an increased current will more and more reduce the benefit of the additional force that was needed for acceleration. This has led to clever combinations with different actuator types, like the reluctance actuator that were not considered suitable before. With the help of fast modern control algorithms and adequate sensors it is possible to compensate partly the negative properties. In the machine tool industry, the actively controlled cutting systems have to handle large forces at moderate speeds. Initially piezoelectric actuators were applied because of their high stiffness, as will be discussed in the last section of this chapter. More recently however, also the use of electromagnetic actuators of the hybrid type has been investigated, because these combine a high force relative to the consumed electric power, a well controllable linear behaviour and a moderate stiffness. Although the actuation range of a hybrid actuator is in the order of one millimetre or less, this is still always much larger than



**Figure 5.39:** Dimensions of the coil, used for the comparison of the three different actuator types.

the stroke of a piezoelectric actuator, with its maximum range of about 1  $\mu\text{m}$  per mm actuator length.

#### 5.4.2.1 Standard coil dimension for the comparison

To illustrate the different properties of the three discussed actuator types, an example with practical values will be presented to close off this section on electromagnetic actuators. As said, several parameters have to be determined, of which the most important are:

- Current to force ratio in relation to power dissipation
- Electrical time constant
- Moving mass

The main force properties of the Lorentz, reluctance and hybrid actuator types, will be compared by means of a small calculation example, all with the same electrical power. For comparison, a coil is chosen with a fixed configuration and total cross section of the windings as shown in Figure 5.39.

As a next step the number of windings for this design needs to be determined. It was already shown in Equation (5.112) that the number of windings has no influence on the dynamic properties of the actuator. The winding volume is fixed, which means that also the mass is constant. Also the power dissipation  $I^2R$  is no factor in the design decision as the current is inversely proportional to  $n$  and the resistance is proportional to  $n^2$ . This means that the only item that determines the number of windings is the electrical power source. More windings imply a higher voltage and a lower current, which means that the number of windings are adapted to the amplifier. In motion systems with very high power also the maximum wire thickness of the supply wires is a

constraint, but for this comparison study this plays no role. This coil has an average length per winding  $\ell_w$  of  $\sim 100$  mm and a cross section  $A_c$  of 50 mm $^2$ . In order to simplify the example, a look up table as shown in Table 5.2 is used for the wiring properties. A practical value of wire is chosen of 0.75 mm, including insulation, enabling easy calculation for this example case. With this wire diameter, the volume can be filled with approximately 100 windings, giving a total length  $\ell_{w,t}$  of ten metres, which results in a resistance of around 0.5 Ohm. One Ampère of current will then give half a Watt of Power. With this coil the performance of the three different actuators will be evaluated.

**Table 5.2:** Look up table for electrical coil windings of copper.

Wire diameter insulated (mm)	Wire diameter (mm)	Windings per cm $^2$	Ohm per 100 m
0.259	0.2	1890	55.8
0.282	0.22	1540	64.1
0.316	0.25	1230	35.7
0.342	0.27	1060	30.6
0.35	0.28	1000	28.5
0.374	0.3	890	24.8
0.396	0.32	750	21.8
0.43	0.35	640	18.2
0.46	0.38	560	15.5
0.487	0.4	510	13.9
0.54	0.45	400	11.2
0.595	0.5	310	8.9
0.65	0.55	270	7.38
0.7	0.6	230	6.21
0.75	0.65	199	5.29
0.81	0.7	174	4.56
0.86	0.75	132	3.97
0.92	0.8	118	3.49
0.97	0.85	106	3.11
1.03	0.9	96	2.76

### 5.4.2.2 Force of the Lorentz actuator

With the Lorentz actuator, the diameter of the chosen coil limits the diameter of the yoke. This also limits the capability of this yoke to transfer a high amount of magnetic flux. For this reason the chosen coil dimensions are more wide than long. In order to avoid saturation of the yoke, the flux density should remain below 2 T. With this value the surface of the air-gap times the flux density in the air-gap should be lower than the maximum flux density times the surface of the yoke. With a small calculation this means that the uniform magnetic field in the air-gap can be maximum 0.7 T. The height of the air-gap is equal to the coil and in the mid position  $F = BI\ell_t$  results in a current to force ratio of 7 N/A. In reality this value will be lower because of stray flux.

### 5.4.2.3 Force of the reluctance actuator

When the same coil is applied in a reluctance actuator Equation (5.80) can be used:

$$F = - \left( \frac{nI}{\ell_g} \right)^2 \frac{\mu_0 A_g}{4} \quad (5.119)$$

If the length of the air-gap equals  $\ell_g = 10^{-3}$  m, which is quite large, and  $A_g = 6 \cdot 10^{-4}$  m<sup>2</sup>, then the resulting force to current ratio would amount up to only 3.75 N/A, which is not yet impressive. When the air-gap is reduced to 0.3 mm, the force with 1 A current already becomes 37.5 N. At full closure of the gap the iron might ultimately saturate with a flux density of around 2 Tesla. With Equation (5.85) the maximum attainable force would be equal to:

$$F_m = - \frac{B_g^2 A_g}{\mu_0} = - \frac{4 \cdot 6 \cdot 10^{-4}}{4\pi \cdot 10^{-7}} \approx 2000 \quad [\text{N}] \quad (5.120)$$

This clearly shows the value of a reluctance actuator but it is necessary to admit that a large error is made due to the approximation on the permeability of the iron ( $\mu_r = \infty$ ). This is far too optimistic when the flux density approaches saturation. By using Hopkinson's law of magnetics, the minimum permeability to realise this value can be determined with the stated flux density of 2 T, using the length  $\ell_y$  of the flux path through the ferromagnetic part and assuming a fully closed airgap  $\ell_g = 0$ :

$$B = \frac{\Phi}{A_g} = \frac{nI}{A_g \mathfrak{R}} = \frac{\mu_0 \mu_r n I}{\ell_y} = 2 \quad [\text{T}] \quad (5.121)$$

With the known values for this example,  $nI = 100$  and  $\ell_y \approx 0.1$  m, the required minimum relative permeability becomes:

$$\mu_r \geq \frac{2\ell_y}{\mu_0 n I} \approx 1.6 \cdot 10^3 \quad (5.122)$$

This high value is not realistic with normal ferromagnetic materials. A value for  $\mu_r$  of 500 is more normal, giving a maximum flux density of about 0.7 T, with a corresponding force of  $\approx 230$  N. Even with this limitation the reluctance actuator still is capable to deliver the largest force of the three electromagnetic actuators.

#### 5.4.2.4 Force of the hybrid actuator

For the force of the hybrid actuator Equation (5.103) is used:

$$F \approx 0.8nI \frac{B_r A_m}{\ell_{g,1} + \ell_{g,2}} \quad (5.123)$$

The same key dimensions as with the other actuators are chosen:

$$A_m = A_g = 6 \cdot 10^{-4} \text{ [m}^2\text{]} \text{ (fits inside the coil)}$$

$$B_r = 1.2 \text{ [T]}$$

$n = 50$  (The coil is split in two to keep the power identical)

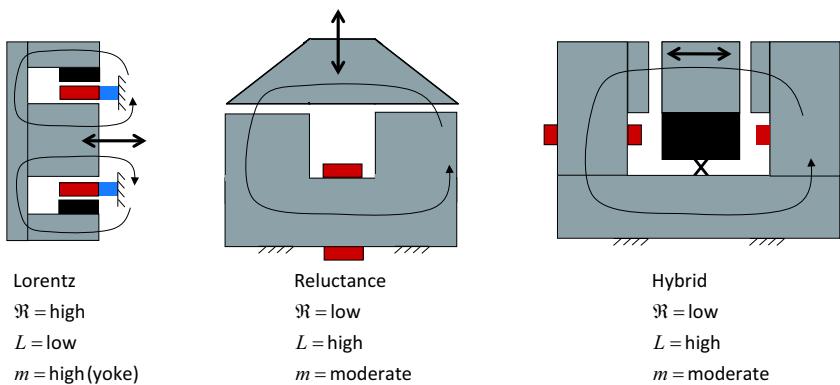
$$\ell_{g,1} + \ell_{g,2} = 1 \cdot 10^{-3} \text{ [m]} \text{ (same stroke as reluctance)}$$

These values result in a linear! current to force ratio of about 30 N/A, which clearly shows the combined performance of “the best of both worlds”, a high force with a linear force to current relation over a moderate range.

#### 5.4.2.5 Dynamic differences

The next criterion for the right actuator choice is the ability to change the force rapidly. This is reflected by the electrical time constant and the related necessary limitation of the jerk. For the same three actuators with the example coil as shown in Figure 5.40, the electrical time constant is purely determined by the reluctance of the flux resulting from the current in the coil. The higher the reluctance, the better it is.

When drawing the field lines of the magnetic field, caused by the current through the coil, in the three actuators, it is clear that the reluctance is quite high for the Lorentz actuator because it includes a large air path. This would even be better when no ferromagnetic part is applied. The reluctance



**Figure 5.40:** Difference in reluctance, self-inductance and mass of the three actuator types scaled to the same coil dimensions.

of the Lorentz actuator is in any case considerably smaller than with the Reluctance and hybrid actuator because of their small air-gap. For this reason, a Lorentz actuator is better suited for high speed precision actuation with fast changing currents, while the other actuators can be applied where high forces in a semi static situation are required. Nevertheless, as mentioned in the previous section, the better force to current ratio of the reluctance and hybrid actuator compensates the negative aspect of the higher self-inductance to some extent because of the lower required current and especially the hybrid actuator is very promising for future applications. A detailed FEM analysis should determine the real optimum as a smaller gap influences the dynamic properties of the hybrid and reluctance actuator in a clearly interrelated way.

#### 5.4.2.6 Moving mass

The last difference is the moving mass. As Figure 5.40 has been scaled to the same coil size, it might be concluded, that the Lorentz actuator is at an advantage when the coil would be the moving part. This however is not always the case. In a precision system the wires to transport the current to the coil determine a relatively high stiffness, that is not preferred for the reason of transmissibility of external vibrations. This can be avoided by choosing the magnetic circuit as the moving part. Furthermore, due to the lower force to power ratio, more current is needed for the same force, which implies a higher power loss. In high power precision systems, the resulting heat has to be taken away, sometimes even by water cooling, before

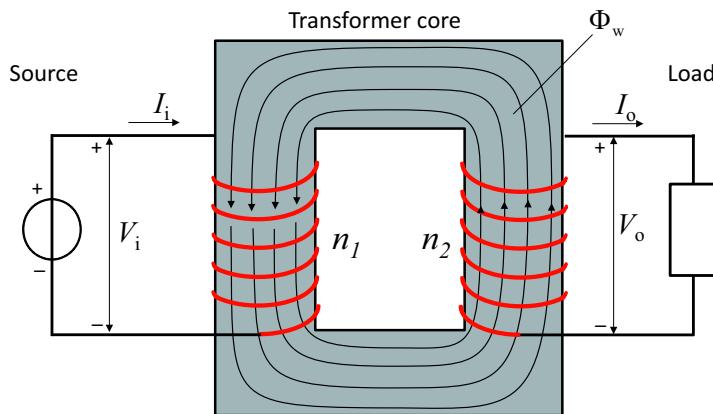
it radiates to other sensitive parts. The necessary cables and hoses and cooling plates considerably increase the stiffness of connection to the coil section and necessitate to take the permanent magnet part as mover.

In fact the Lorentz actuator has the largest moving mass of all, when looking only to the force related to the power. Furthermore, when considering that the variable reluctance- and hybrid actuator both have the possibility to add more coil windings, it is clear that these actuators are absolutely preferred from a high force to power ratio point of view in applications where the dynamic properties and linearity are less important.

## 5.5 Intermezzo: electric transformers

As a transition marker between the electromagnetic actuators and Chapter 6 on electronics, the operation principle of an electric transformer is a very valid candidate. This important component is mostly called just shortly a transformer, because it is in most cases clear what is meant based on the context. An electric transformer is a passive component that transforms an alternating input voltage and current into an alternating output voltage and current, often with a different magnitude but without changing the power. This means that a high input voltage with a low input current will be transformed into a low output voltage with a high output current or the other way around. A transformer only changes the ratio between the voltage and current of an alternating signal and this transformation takes place in two steps. First the electric power is transformed into magnetic power in a similar manner as with reluctance actuators, but then, instead of transforming the magnetic power into motion, the transformation is reversed again into electricity at another set of terminals, without a conductive connection with the source. This principle gives an inherent galvanic insulation, which is one of the main reasons of the application of a transformer. In mechatronic systems the transformer is mainly used in power supplies, but it is also used for measuring signals at long distances, where galvanic insulation prevents interference by ground loops, as will be presented in Chapter 8.

Figure 5.41 shows the layout of a basic transformer with two coils wound on a shared ferromagnetic yoke, called the *transformer core*. Because of the low reluctance of the ferromagnetic material both coils share the same flux. In principle a multitude of different coils could be applied, either connected in series, like in the *autotransformer* or all kept separately to create many different voltages that are mutually galvanic insulated. To explain the principle, the shown configuration with only two separate coils is sufficient.



**Figure 5.41:** In an electric transformer two coils are wound around the same ferromagnetic yoke, called the transformer core. Because they share the same flux, the voltage of both coils relate to the ratio between the windings of each coil.

In order to transform the electric current and voltage to another value, the number of windings per coil is different and a higher number of windings corresponds to a higher voltage with a lower current level. The *primary windings* are connected to the electrical source and the *secondary windings* are connected to the load. In principle a transformer can transform electric energy in two directions and as such the term primary and secondary windings is not logical, but this distinction has been made for practical reasons. It is for instance not advisable to reverse the connections of a transformer that is designed to transform the dangerous mains voltage into a safe lower voltage, as the result would be an even higher voltage and eventually the transformer will burn out when the fuses are not blown before.

The figure also illustrates the galvanic insulation between the primary and secondary windings as they each consist of insulated wires.

### 5.5.1 Ideal transformer

In an ideal transformer, the windings are assumed to have no resistance and the ferromagnetic material is assumed to have infinite permeability and no magnetic limitations nor energy losses with changing magnetic fields.

To explain the working principle the relation between the magnetic field and the alternating input voltage  $V_i$  over the primary winding  $n_1$  of an ideal

transformer is determined by the same relation as with the self-inductance of a coil:

$$V_i(\omega) = n_1 \frac{d\Phi_w}{dt} = \hat{V}_i \sin \omega t \quad (5.124)$$

where  $\Phi_w$  is the flux in the core.

The secondary winding shares the same flux which means that the output voltage  $V_o$  equals:

$$V_o(\omega) = n_2 \frac{d\Phi_w}{dt} = \hat{V}_o \sin \omega t \quad (5.125)$$

With this equation the voltages from the primary and secondary windings appear to relate as follows:

$$\frac{V_o}{V_i} = \frac{n_2}{n_1} \quad (5.126)$$

To determine the primary and secondary currents, the flux in the core is first calculated. From Equation (5.124) the change of flux in the core should equal:

$$\frac{d\Phi_w}{dt} = \frac{\hat{V}_i}{n_1} \sin \omega t \quad (5.127)$$

By integrating over time, the flux becomes:

$$\Phi_w(\omega) = -\frac{\hat{V}_i}{\omega n_1} \cos \omega t \quad (5.128)$$

The primary current can be calculated from the relation between flux and current, corresponding with Hopkinson's law of magnetics:

$$\Phi_w = \frac{n_1 I_i}{R} \quad (5.129)$$

In the situation without an external load the secondary current is zero and the unloaded primary current  $I_{i,0}$  equals

$$I_{i,0}(\omega) = \frac{\Phi_w(\omega) R}{n_1} = -\frac{\hat{V}_i R}{\omega n_1^2} \cos \omega t \quad (5.130)$$

A low reluctance of the core results in a low primary current in the unloaded situation. In an ideal transformer with  $\mu_r$  is infinite,  $I_{i,0}$  will be equal to zero. As soon as a load is present at the secondary windings, the resulting secondary current  $I_o$  creates an magnetic flux  $\Phi_{L,o}$  in the core that adds to  $\Phi_w$ . A primary current will start to flow with a corresponding flux  $\Phi_{L,i}$  that compensates the additional flux caused by the secondary current, because

the total flux has to remain unchanged, corresponding with the primary voltage. In simplified form without  $\omega t$  this reads as follows for an ideal transformer:

$$\Delta\Phi_w = \Phi_{I,i} + \Phi_{I,o} = 0 \quad (5.131)$$

The directions of the voltages and currents as defined in Figure 5.41 follow from Ampère's and Faraday's laws. The flux from the output current should point into the other direction of the flux by the primary current. With the defined directions the equation becomes:

$$\Phi_{I,i} + \Phi_{I,o} = \frac{n_1 I_i - n_2 I_o}{R} = 0 \implies n_1 I_i = n_2 I_o \quad (5.132)$$

As a consequence the currents from the primary and secondary windings relate as follows:

$$\frac{I_o}{I_i} = \frac{n_1}{n_2} \quad (5.133)$$

The result can be verified by a sanity check, based on the fact that an ideal transformer does not dissipate power. This means:

$$P = P_i = I_i V_i = P_o = I_o V_o \quad (5.134)$$

Which coincides with the found relations.

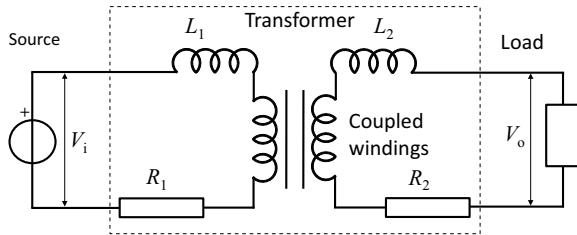
### 5.5.2 Real transformer

In reality a transformer is not ideal as the windings have a certain resistance and the core is not ideally magnetically conducting. This means in the first place, that the coupling of the flux of the primary and secondary windings is not 100 %. To improve this coupling, the primary and secondary windings are preferably wound as close as possible together. To illustrate the effect of this, the electrical equivalent of a real transformer is shown in Figure 5.42 with the series resistance and uncoupled self-inductance in the primary and secondary windings. The resistance results in power loss and a decreased output voltage in case of a loading current. The uncoupled self-inductance will influence the dynamic behaviour of the system.

The most important limitation of a transformer is however determined by the magnetic saturation of the core.

With Equation (5.128) the maximum flux and flux density in a transformer can be derived:

$$|\hat{\Phi}_w| = |\hat{B} A_c| = \left| \frac{\hat{V}_p}{\omega n_1} \right| \quad (5.135)$$



**Figure 5.42:** Electrical equivalent of a real transformer.

Because the flux density has to remain below the saturation level  $B_s$ , the cross section of the core needs to be larger than:

$$A_c \geq \frac{\hat{V}_p}{B_s \omega n_1} \quad (5.136)$$

The conclusion that can be drawn from this equation is that a transformer can not be used to transform DC voltages as that would require an infinitely large core. For AC voltages of low frequencies the number of the primary windings could be increased, but that will automatically lead to an increase in the resistive losses. This means that the design of a transformer is a compromise between power, efficiency and size.

Because of the relation with the frequency it is unavoidable that high power transformers that have to operate at 50 Hz are very heavy. To solve that problem, the *switched-mode power supply* has been created. In this principle first the low-frequency mains voltage is converted electronically into a very high frequency of several kHz, before it is transformed to a lower voltage by means of a small high-frequency transformer. The operation principle is comparable with the switched-mode power amplifiers, that will be presented in Section 6.3.3 in Chapter 6.

## 5.6 Piezoelectric actuators

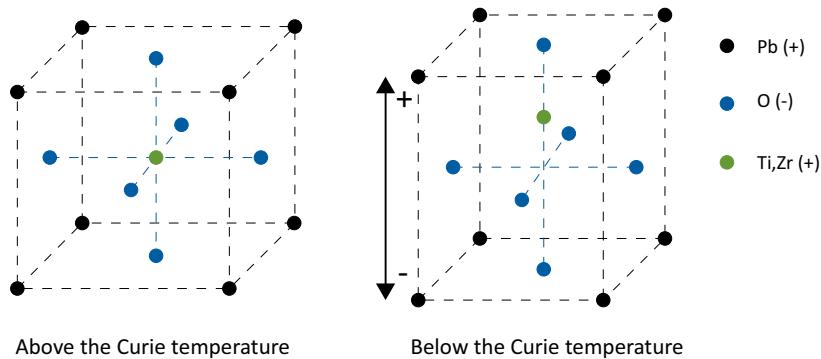
### 5.6.1 Piezoelectricity

The word “piezo” is derived from the Greek word “πίεζω”, which means “to press”. The piezoelectric effect was first observed by the Pierre Curie together with his brother Paul-Jacques Curie (1856 – 1941) who also was a physicist. They discovered that under compression a quartz crystal is generating an electric potential, which is called the *direct piezoelectric effect*. It later has been discovered that this effect also is reversible, which is called the *converse piezoelectric effect* or *inverse piezoelectric effect*: when an electric field is applied to the piezoelectric material by applying a voltage to the electrodes at the piezo-surface, the material expands or contracts depending on the direction of the electric field.

Besides quartz, several other materials are found to show a piezoelectric behaviour:

- Some natural crystals like Rochelle salt.
- Some natural tissue such as bones and wood.
- Synthetic piezo-ceramics such as Lead Zirconate Titanate (PZT) and Lead Lanthanum Zirconate Titanate (PLZT).
- Synthetic polymers such as Polyvinylidene Fluoride (PVDF).

The reason for the piezoelectric behaviour is an asymmetry in the molecular structure of the piezoelectric material which gives an asymmetry in the charge distribution inside the materials as shown in Figure 5.43. This charge asymmetry is called *polarisation* and is caused by large atoms that are located inside a crystal structure formed by other atoms. The PZT material that is shown in the figure as example consists of a Lead oxide ( $\text{PbO}_3$ ) crystalline structure where a certain amount of Titanium (Ti) and Zirconium (Zr) atoms are located according to a *Perovskite structure*. At elevated temperatures the structure is large enough to accommodate the Ti/Zr atoms. In that case the material is isotropic and does not possess piezoelectric properties. When cooling down below the *Curie temperature* the shrinking structure is deformed into a tetragonal ferroelectric structure where the Ti/Zr atoms are “pushed” aside with a simultaneous displacement of the oxygen atoms. The oxygen atoms are negative charged as they received electrons from the other atoms by the chemical reaction that created the



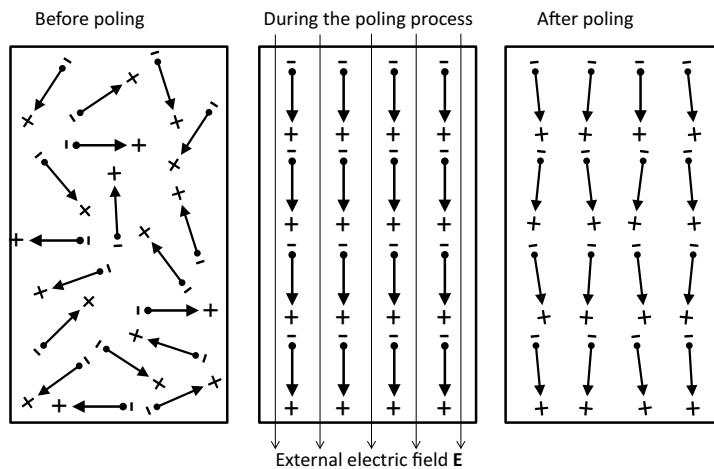
**Figure 5.43:** The piezoelectric effect in PZT material is caused by an asymmetry in the crystal structure of the material where large atoms like Titanium or Zirconium are “frozen” into the structure when cooling down from above the Curie temperature.

structure which means that a charge displacement occurs, creating a dipole in the crystal.

### 5.6.1.1 Poling

Similar to a permanent magnet material, piezoelectric material with the same polarisation direction is grouped in Weiss domains. In a multi-crystalline material these domains are normally randomly oriented and the piezoelectric effect is not observed on a macroscopic level. To utilise the piezoelectric properties on the macroscopic scale, the Weiss domains have to be aligned. In some materials like quartz this happens naturally, but like in PZT a poling process is necessary for this alignment, as is shown in Figure 5.44. During poling a strong electric field ( $>2\text{ kV/mm}$ ) is applied to the PZT material while its temperature gets elevated close to the Curie temperature which enables the Weiss domains to align to the external electric field. When cooling down the PZT material and switching off the external field, the domains rotate slightly such that the material becomes neutrally charged again, but the main orientation is maintained and the material has obtained a remnant polarisation.

After poling, the piezoelectric properties are also observed on the macroscopic scale. An external mechanical stress or electric field can disturb the macroscopic balance of the piezo-material with observable effects on the outside of the material. In Figure 5.45 on the left the direct piezoelectric effect is shown, where an external force on the piezo-material generates a



**Figure 5.44:** Poling process of a piezoelectric material. Without poling the Weiss domains are not aligned. During poling a strong external electric field is applied and also the temperature is elevated to the Curie temperature. After cooling down the material, the electric field is switched off and the material is “frozen” into a state where the piezoelectric effect is present on the macroscopic scale as the Weiss domains stay almost perfectly aligned.

voltage at the electrodes of the piezo. The right side of the figure illustrates the converse piezoelectric effect where an external voltage, generating an electric field across the piezo-material, causes the piezo-material to expand or contract, depending on the direction of the voltage.

## 5.6.2 Transducer models

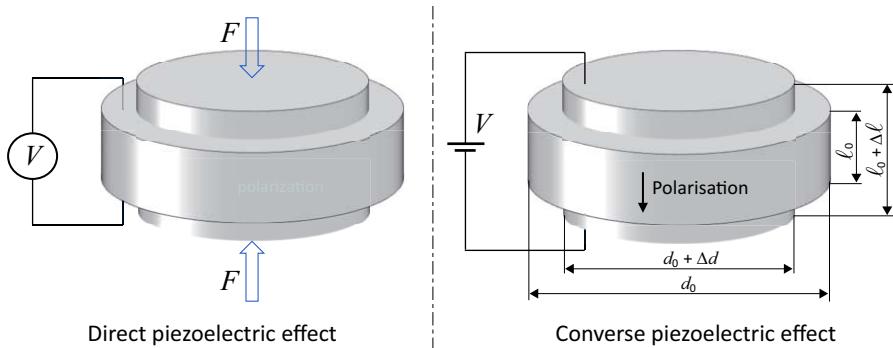
Piezoelectricity is the combination of the electric and mechanical behavior of the material, which is given by the equation for the spatial electric displacement  $\mathbf{D}$ :

$$\mathbf{D} = \epsilon \mathbf{E}, \quad (5.137)$$

with permittivity  $\epsilon$ , and the electric field  $\mathbf{E}$ ,

The Hooke – Newton law, as was defined in Chapter 3, relates the strain  $\mathbf{S}$  of a material to its compliance  $\mathbf{C}$  and the externally applied stress  $\mathbf{T}$ , when written vectorial in multi dimensions:

$$\mathbf{S} = \mathbf{CT} \quad (5.138)$$



**Figure 5.45:** Schematic of the piezoelectric effect. Under compression a voltage can be measured at the electrodes (direct piezoelectric effect). When applying a voltage to the electrodes, depending on the sign, the piezo-material expands or contracts (converse piezoelectric effect).

Combining these for all directions of the material into the so-called coupled equations, according to the IEEE standard on piezoelectricity, the constitutional laws for a piezo<sup>4</sup>-material in strain charge format are:

$$\mathbf{S} = \mathbf{C}_E \mathbf{T} + \mathbf{d}^T \mathbf{E} \quad (5.139)$$

$$\mathbf{D} = \mathbf{d} \mathbf{T} + \boldsymbol{\varepsilon}_T \mathbf{E}, \quad (5.140)$$

with the following variables, listed with the corresponding units and dimensions:

$\mathbf{S}$  in [-]  $6 \times 1$  strain vector

$\mathbf{T}$  in [ $\text{N}/\text{m}^2$ ]  $6 \times 1$  stress vector

$\mathbf{E}$  in [ $\text{V}/\text{m}$ ]  $3 \times 1$  electric field vector

$\mathbf{D}$  in [ $\text{C}/\text{m}^2$ ]  $3 \times 1$  electric displacement vector,

and the following constants:

$\mathbf{C}_E$  in [ $\text{m}^2/\text{N}$ ]  $6 \times 6$  compliance matrix with constant electrical field

$\mathbf{d}$  in [ $\text{m}/\text{V}$ ] or [ $\text{C}/\text{N}$ ]  $3 \times 6$  piezoelectric coefficient matrix

$\boldsymbol{\varepsilon}_T$  in [ $\text{F}/\text{m}$ ]  $3 \times 3$  dielectric coefficient matrix with constant stress.

<sup>4</sup>Often the term piezoelectric is just shortened into “piezo”.

In data sheets of piezo-actuators, often the piezoelectric coefficient  $d_{ij}$  is given, where the indices  $i$  and  $j$  indicate the directions of polarisation and strain. So  $d_{33}$  indicates the strain parallel to the polarisation and is referred to as the *piezo-gain* for stack actuators, while  $d_{31}$  indicates that the strain is orthogonal to the polarisation giving the piezo-gain for instance for tube actuators.

The constitutional laws describe the physical behaviour of a piezoelectric element. For a better understanding of how this material works as a macroscopic actuator, such as a piezoelectric stack-actuator, these equations can be converted from the stress and electric field strength into a format that used the external force  $F$  in [N] and applied voltage  $V$  in [V], and from strain and electric displacement into displacement  $\Delta\ell$  in [m] and charge  $q$  in [C]. For simplicity and also because it is close to most practical applications for piezoelectric actuators, a homogenous material, electric field and stress is assumed, as well as their respective alignment with the polarisation and strain. Under these conditions, this conversion gives the following set of simple scalar multiplications according to the geometry of the actuator with cross section  $A$  and thickness  $\ell_0$ , resulting in:

$$\begin{aligned} F &= T \cdot A \\ V &= E \cdot \ell_0 \\ \Delta\ell &= S \cdot \ell_0 \\ q &= D \cdot A, \end{aligned} \tag{5.141}$$

The material equations for the stiffness  $k_{pz}$  in [N/m] and capacitance  $C$  in [F] are as follows:

$$\begin{aligned} k_{pz} &= \frac{A}{s_E \cdot \ell_0} \\ C &= \frac{\varepsilon_T \cdot A}{\ell_0}. \end{aligned} \tag{5.142}$$

When using these results in the constitutional laws as defined in Equation (5.139) the following simple equations are derived:

$$\begin{aligned} \Delta\ell(t) &= k_{pz}^{-1} \cdot F(t) + d_{ij} \cdot V(t) \\ q(t) &= d_{ij} \cdot F(t) + C \cdot V(t). \end{aligned} \tag{5.143}$$

From these equations it can be concluded that an external voltage, applied to the piezo-actuator, charges the capacitance of the piezo with a charge  $q$

and that it causes a displacement  $\Delta\ell$ . Simultaneously an external force on the piezo-actuator is manifested by a displacement  $\Delta\ell$  due to the high but finite stiffness of the piezo-material and that it generates a charge  $q$  as well. These properties of piezoelectric materials enable these electro-mechanical actuators to be used for both actuation as well as sensing applications, as described before.

The transducer model, that is given in this section, describes a linear mathematical model of the interplay between the electrical and the mechanical domain. Although this model is very important to describe piezoelectric actuators, for analysis as well as system design, it does not cover the non-linear properties of piezoelectric materials, which are discussed in the following section.

### 5.6.3 Nonlinearity of piezoelectric transducers

Although piezoelectric actuators have no moving parts where friction and backlash can occur, which might otherwise compromise the precision of the actuator such as in a DC-motor with a gear box, these actuators show other non-linear behavior such as creep and hysteresis.

#### 5.6.3.1 Creep

A piezoelectric material shows *creep* after changes of the applied voltage. This effect is observed as a slow drift, caused by the effect of the actuation voltage on the remnant polarisation of the piezoelectric material. This creep effect is logarithmic as function of time according to the following equation:

$$\Delta\ell_{(t)} = \Delta\ell_{(t=0.1)} \cdot (1 + \gamma \cdot \lg \frac{t}{0.1}), \quad (5.144)$$

where  $\Delta\ell_{(t=0.1)}$  is the displacement 0.1 seconds after application of the change in the actuation voltage when the fast dynamic transients have settled and  $\gamma$  is the creep constant, which typically is in the order of 1 – 2 %.

After a few hours, the displacement due to creep can be as large as 10 % of the total displacement, which may lead to significant positioning errors in open loop operated positioning systems that are based on piezoelectric actuation. If the coefficients connected to piezo-creep are known well enough, it can be compensated (or at least significantly reduced) in an open loop manner without the use of position sensors. Even with this compensation already a small uncertainty in these parameters may accumulate in a positioning error that is not acceptable in precision positioning applications that need

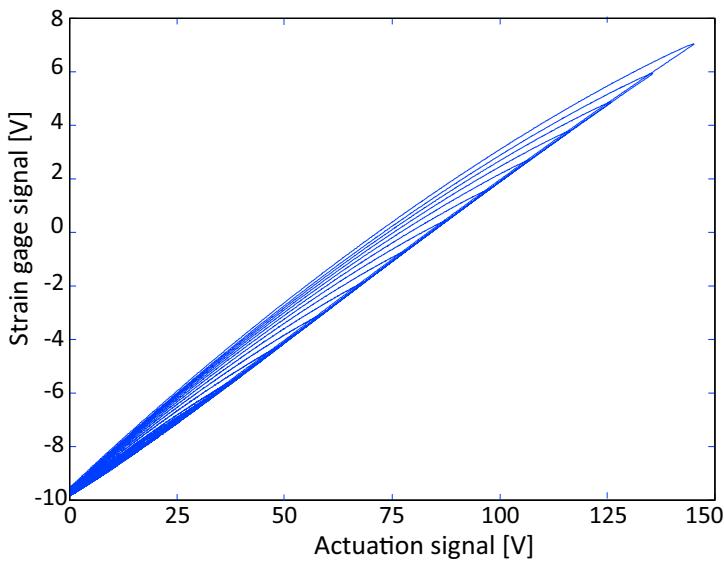
a long term stability in the order of nanometres or even less. A position sensor in combination with the piezoelectric actuator can be an alternative solution while a very small integrating action will fully compensate the positioning error due to the slow creep of the piezoelectric material.

### 5.6.3.2 Hysteresis

*Hysteresis* is a non-linearity that occurs in piezoelectric material and is manifested by a difference between the motion path for a raising actuation voltage as compared to a declining actuation voltage, and can be as large as 15 % of the desired elongation of the piezo-actuator. The source for this hysteresis is molecular friction, non-ideal material properties and polarisation. Hysteresis can be described in a way that some Weiss domains align easier to the external field than others, which causes them to align earlier on the raising actuation voltage as well as on the declining actuation voltage as compared to those domains who align not as easy, resulting in the difference between the raising and falling actuator elongation as a function of the actuation voltage. This means that the actual elongation of a piezo-actuator depends both on the currently applied voltage as well as on the last turning point where the time derivative of the actuation voltage changed its sign. A demonstration of various hysteresis loops is shown in Figure 5.46, where a triangular actuation voltage with its amplitude declining over time has been applied to a piezoelectric stack actuator, while the elongation of the actuator has been measured by a strain gage sensor.

It is worthwhile to mention here that the hysteresis effect is occurring mainly in voltage driven piezo-actuation. When the driving signal is controlled with respect to the charge  $q$  that is applied to the piezoelectric actuator, hysteresis does not occur. This can be explained from the following. The converse piezoelectric effect corresponds with a charge displacement that creates the deformation of the crystal while this charge displacement relates to the voltage by the capacitance of the actuator. The main part of the hysteresis occurs in the dielectric properties of the material that determine the capacitance value. When the charge is directly controlled, this dielectric value has no influence anymore. In Chapter 8 a similar approach will be presented to measure acceleration by charge amplification when using the direct piezoelectric effect.

For voltage controlled systems, piezo-hysteresis can be modelled by so-called *hysteresis operators*, such as a *Preisach model*, and can be compensated for in an open-loop manner. However, just as in the case of creep, several



**Figure 5.46:** Measured hysteresis loops of a piezoelectric stack actuator for various amplitudes of the input signal over a 150 V range. The strain gauge signal is proportional to the measured elongation of the actuator.

parameters of the hysteresis model have to be known perfectly, including the actuation history, in order to achieve a precise compensation of the hysteresis.

For applications where accuracy is not as important as precise positioning in a repetitive fashion, such as the scanning motion in an Atomic Force Microscope, hysteresis is not much of a problem and calibration and pre-shaping of the driving voltage allows to compensate for most of the hysteresis. In these applications the repeatability of the position (precision) can easily be made better than 1 nm if only the rising **or** the falling voltage branch is used, even though the difference between these two branches may well be in the order of a micrometre.

Another way to compensate for piezo-hysteresis is feedback control, just in the same way as described for the compensation of creep.

### 5.6.3.3 Aging

Another effect that occurs in piezoelectric material is *aging*, where due to de-poling the piezo-gain reduces over time. This phenomenon happens on the scale of several months or years. When the piezoelectric material is

used as an actuator, the material gets re-polarised whenever a high field gets applied, which means close to the maximum drive voltage. In this case material aging can be neglected. When the piezoelectric material is used as a sensor or for the generation of charge, where no external voltage is applied, reduction of the piezo-gain due to aging may have to be considered.

### 5.6.4 Mechanical considerations

In contrast to Lorentz actuators, which have zero stiffness properties, most piezoelectric actuators have a high stiffness due to the fact that they are made of ceramic materials. The mechanical properties of the piezo-actuator have to be taken into account when designing a mechatronic positioning system.

The high stiffness properties of piezo-actuators enable to realise very fast positioning systems with high actuation forces in the range of several kN and very high resonance frequencies  $f_0$ , which is next to the sub-nanometre resolution of piezo-actuators one of the main advantages to use these types of actuators. One disadvantage of these actuators that also is linked to the property of high stiffness is the limited actuation range, which as a rule of thumb is that a piezo-ceramic can expand by about 0.1 % of its total length at the maximum applied voltage.

Another possible disadvantage of the high stiffness can be the high transmission of vibrations from the support. This makes this actuator less suitable for applications like the optical pick-up unit of a CD player or stages in a wafer scanner where disturbances from the vibrating environment have to be avoided. Piezo-electric actuators are best applied when these vibrations from the surrounding can be sufficiently reduced by additional methods or when the total system is so small and integrated that accelerations by the overall movement of the total system can be neglected.

#### 5.6.4.1 Piezo-stiffness

In the data sheet of each actuator some typical parameters are specified. Next to the geometric properties and maximum actuation voltage, important parameters for the mechanical design are the blocking force  $F_{max}$  and the displacement range for the unloaded and unrestrained actuator  $\Delta\ell_0$ . Often also the actuator stiffness is given, denoting the small signal stiffness  $k_{pz}$ , relating to the other parameters via

$$F_{max} \approx k_{pz} \cdot \Delta\ell_0. \quad (5.145)$$

The blocking force  $F_{\max}$  denotes the force that a piezo-actuator generates when it is confined on both sides by an infinite stiff support. This means that the entire force generated by the converse piezoelectric effect presses against the clamping support structure as no elongation of the piezo occurs. When the piezo-actuator is not confined or restrained, the force generated by the actuator is available for elongation of the piezo, which means that the actuation force is in balance with the intrinsic spring stiffness of the piezo-actuator, resulting in a displacement  $\Delta\ell_0$  as shown in Figure 5.45. The given value of  $k_{pz}$  is only valid for small signal conditions. For larger signal amplitudes, another term is superimposed on  $k_{pz}$  that is linked to polarisation effects of the material, where the large signal stiffness can be up to a factor of two smaller than the small signal stiffness. Furthermore the output impedance of the power amplifier driving the piezo-actuator has an influence on the actuator stiffness, as the charge that is built up in the material due to external forces, if not drained out of the piezo-material, generates a counter-force. This is the reason why a piezo-actuator with open electrodes appears stiffer than a piezo-actuator with shortened electrodes .

#### 5.6.4.2 Actuator types

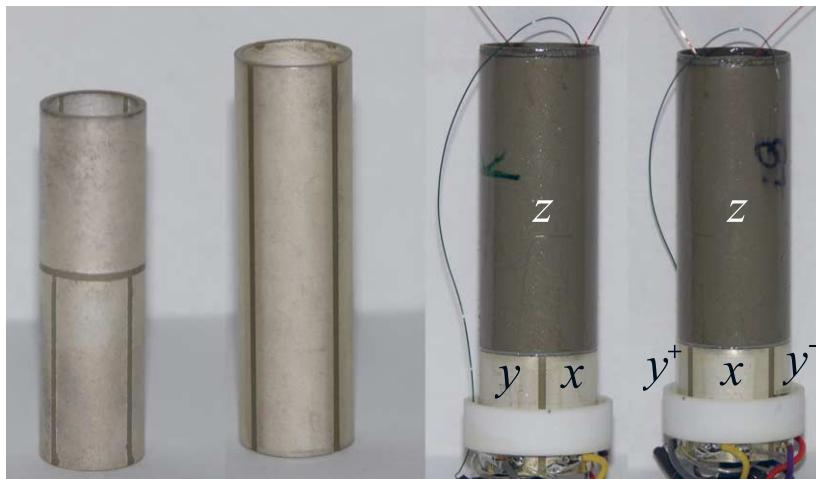
Many different types and shapes of piezoelectric actuators exist, such as piezoelectric bending elements, tubes, cones and plates, shear-actuators and stack-actuators. Only two of the most important versions are presented in some more detail because the presented design considerations are equally valid for the other actuator types.

*Piezoelectric tube actuators* enable positioning in all three linear spatial directions by means of only one single actuator with multiple electrodes on its barrel as shown in Figure 5.47. The lateral motion in the  $x$ - and  $y$ -directions is generated by elongating the piezo-material at one electrode of the tube while contracting it at the electrode on the opposite side of the tube where the same signal gets applied but with a reversed sign of the actuation voltage . This causes a bending of the tube in  $x$ - or  $y$ -direction where the displacement  $\Delta x$  amounts to

$$\Delta x = \frac{2\sqrt{2} \cdot d_{31} \cdot \ell^2 \cdot V}{\pi \cdot D_i \cdot d}, \quad (5.146)$$

with the tube geometry of length  $\ell$ , wall thickness  $d$ , and inner diameter  $D_i$ , and actuation voltage  $v$ .

Actuation in  $z$ -direction is done by applying a voltage on all electrodes on the entire circumference causing a uniform elongation with a vertical



**Figure 5.47:** A tube-scanner consists of one tubular piezoelectric element with different electrodes at the inside and outside. Different electrode configurations can be used depending on the application. The  $x$ - and  $y$ -actuators of the examples on the right each consist of two oppositely working active areas at each side of the tube. The  $z$ - actuator is a uniform linear expanding element.

displacement  $\Delta z$  of

$$\Delta z = d_{31} \cdot \ell \cdot \frac{V}{d} \quad (5.147)$$

Piezoelectric tubes are very popular for 3D-positioning applications with high precision and resolution, because of their simple design, low cost and low capacitance, which makes the design of the power amplifier to drive the actuator easier. A disadvantage of tube actuators, however, is that the  $x-y$ -motion is defining a spherical segment that causes a cross-coupling also into the vertical  $z$ -direction, which can be a compromising factor when very precise positioning is required.

*Piezo-stack actuators* as used in the example in Figure 5.48 are multi-layer actuators, typically build from sheets of  $d_{33}$  piezo-material. The individual about  $50\mu\text{m}$ -thick sheets have electrodes on both sides to apply the actuation voltage and are stacked and glued together to form a longer actuator. This stacking of multiple layers allows to generate larger strokes with smaller actuation voltages, in the order of 150 V, as the strength of the electric field is given by the voltage divided by the layer thickness. Without stacking of the piezo-layers, the actuation voltage for the same actuator length would

be in the order of several kV, which is more difficult to achieve with standard electronics and also poses a larger safety risk. As the poling and actuation direction is in the stacking direction, the number of layers  $n$  chosen will define the unloaded length  $\ell_0$  and the positioning range of the stack-actuator.  $\ell_0$  eventually is  $n$  times the unloaded and unactuated thickness of the individual layer  $h$  plus the thickness of the electrodes, glue lines and potentially a layer of unactuated ceramic material at both ends of the stack-actuator for isolation. With the actuation voltage  $V$  the actuator displacement  $\Delta\ell$  amounts to:

$$\Delta\ell = n \cdot h \cdot d_{33} \cdot \frac{V}{h}. \quad (5.148)$$

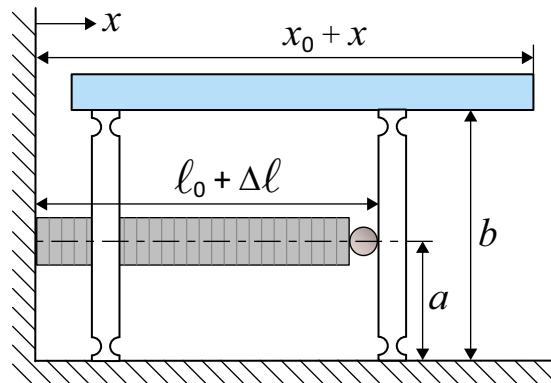
#### 5.6.4.3 Actuator integration

An appropriate integration of the piezo-actuator in the mechanical structure of the mechatronic system allows to prevent failure of the actuator as well as to reduce or amplify the positioning range for actuation.

*Pre-loading* the actuators by an external compressive force enables the prevention of fatal tensional loading forces. Due to their ceramic nature piezo-actuators are very stiff but also very brittle. Furthermore in case of stack-actuators, the individual piezo-sheets are glued together generating a potential weakness at the bond-line. This means that although piezoelectric material can withstand very high compressive forces, they should not be loaded in tension, torsion, or shear if not explicitly specified for this purpose. For applications where the piezoelectric actuator has to position a mass in both directions with pushing and pulling, it must be assured that the acceleration at pulling does not generate too large tension forces. To prevent accidental damage by these tensional forces, piezoelectric actuators are often pre-loaded with springs. This can either be done with a simple helical spring that works in parallel with the piezo-actuator. A better alternative are flexure springs that at the same time also confine the piezo-motion to the direction of actuation and suppress parasitic rocking and torsional motion of the actuator. In both cases the stiffness  $k_f$  of the pre-loading (flexure) spring increases the stiffness of the total system  $k_{\text{tot}}$  to

$$k_{\text{tot}} = k_{\text{pz}} + k_f. \quad (5.149)$$

While the loading mass has no influence on the (static) positioning range of the piezo-actuator, the loading spring reduces the maximum actuation range as the actuation force now has to work against the piezo-stiffness as well as the loading spring. So a loading spring that is as stiff as the



**Figure 5.48:** Amplification of the elongation of a piezo-actuator by a parallel flexure mechanism with amplification factor  $r = b/a$ .

piezo itself  $k_f = k_{pz}$  will reduce the actuation range in half. What may be desirable in some cases is to have a constant pre-loading force that is not a function of the position but ensures that a constant pre-loading force of the actuator is present. This can be achieved by an elongated spring with a low stiffness. Without pre-loading, piezo-actuators can de-laminate when being contracted too quickly, where the high tensile forces can cause micro-cracks or even total fracture of the piezo-material.

#### 5.6.4.4 Mechanical amplification

*Mechanical amplification* by means of leverage enables a larger positioning range than the maximum stroke of the actuator would allow. Figure 5.48 shows a amplification scheme based on parallel flexures that allows to amplify the extension of the piezo-actuator stack  $\Delta\ell$  to the desired displacement  $\Delta x$  with the ratio  $r$  of the total lever length  $b$  to the position  $a$  at the flexure where the piezo-actuator is attached.

$$r = \frac{b}{a} \quad (5.150)$$

$$\Delta x = r \cdot \Delta\ell \quad (5.151)$$

$$(5.152)$$

Due to the mechanical amplification  $r$  the effective stiffness  $k_{tot}$  on the load side as well as the resonance frequency  $f_{res}$ , which depends on the total positioned mass and the effective system stiffness, are reduced according to

the equations

$$k_{\text{tot}} = \frac{k_{\text{pz}}}{r^2} \quad (5.153)$$

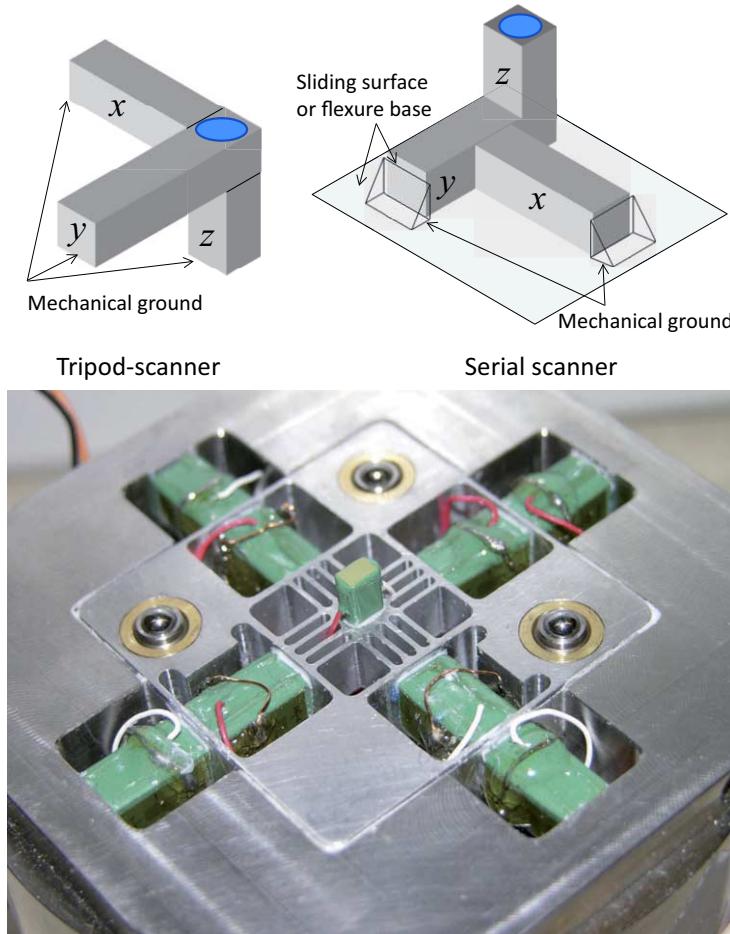
$$f_{\text{res}} = \frac{f_0}{r}. \quad (5.154)$$

#### 5.6.4.5 Multiple directions by stacking

Except for piezoelectric tubes, piezo-actuators are unidirectional actuators. When positioning in more than one direction is required, individual actuators can be combined by means of mechanical design, including flexure mechanisms and stacking of multiple actuators. Figure 5.49 shows three different implementations of a three degree of freedom *piezoelectric scanner* that are based on the stacking of three separate actuators. The typical positioning range of these scanners can vary from a few hundred nanometres up to about 150 micrometres, depending on the length of the applied stack. The first design on top shows a *tripod-scanner* where three stack-actuators are glued together orthogonally, forming a tripod. The advantage of this design is its simplicity and that all three directions are implemented in parallel. A disadvantage of the tripod design is, similar to the piezo-tube, that this design also shows a cross-coupling, called *scanner bow*, between the  $x$ - and  $y$ -direction into the  $z$ -direction, which is caused by the fact that for each actuator the mounting point of the two other respective actuators act as pivoting points.

Two designs that do not suffer from scanner bow are the serial scanner on the right side of the figure and the flexure scanner on the bottom. In the *serial scanner* the individual actuation directions are implemented in a serial way, stacked on top of each other. This means that the  $x$ -actuator moves the  $y$ - and the  $z$ -actuator, the  $y$ -actuator moves the  $z$ -actuator, and only the  $z$ -actuator is moving only the object of interest, shown in the schematic by the small sample disc on top of the  $z$ -actuator. The disadvantage of the serial design is that the speed performance for the lower positioning directions is significantly less than for the ones that are stacked on top, due to the much larger mass that has to be moved.

This difference in performance for the different positioning axes is avoided by implementing the actuation for the various directions in parallel. This is achieved in the high speed *flexure-scanner* for a scanning probe microscope (SPM) as shown at the bottom of the picture. This design shows a parallel implementation of the  $x$ - and  $y$ -directions by combining the individual



**Figure 5.49:** Different designs of piezoelectric scanners, positioning systems for translation in all three spatial directions. The three degrees of freedom are each determined by separate piezoelectric actuators. The high-speed scanner of the picture is able to avoid tensional stresses in the  $x$ - and  $y$ -direction by actuating from two sides.

stack actuators that work in a push-pull configuration by means of a flexure mechanism, resulting in the same performance and resonance frequencies for the  $x$ - and  $y$ -directions. The  $z$ -actuator is located on top of the flexure structure and is implemented in a serial way. In the particular application of the SPM, the highest positioning bandwidth is required in the  $z$ -direction and therefore the total mass moved by the  $z$ -actuator has been minimised.

## 5.6.5 Electrical considerations

As explained in the section on piezoelectricity, poling of the piezo-material is achieved by aligning the Weiss domains with the application of a strong external electric field as shown in Figure 5.44. Actuation of the piezoelectric material is done by applying an electric field according to Equation 5.139. As can be seen from the hysteresis loop in Figure 5.46 only positive actuation voltages are applied. Piezoelectric material is able to accept weak electric fields in the opposite direction to about 10 % of the maximum field strength in the positive actuation direction. A stronger electric field in the opposite direction can cause *de-polarisation* of the material and the remnant macroscopic polarisation in the material due the poling process would be lost. For that reason piezoelectric materials that are not specified for symmetric actuation voltages should only be driven with positive voltages in order to avoid de-polarisation of the material.

### 5.6.5.1 Charge vs. voltage control

As shortly mentioned before when presenting hysteresis, there are two methods to control a piezoelectric actuator. One way is to control the applied voltage and the second is to control the charge that is applied to the piezo-actuator. Both have their advantages and disadvantages.

In case of *charge control* the hysteresis is avoided as the elongation of the actuator is a linear function of the applied charge. To control the charge, the piezo-actuator needs to be driven by a charge source. This means that the current is measured and integrated in the amplifier and actively controlled to the demanded value by means of feedback. With this method the voltage at the output of the amplifier is not observed nor controlled which means that the amplifier has an infinite output impedance. A consequence of this infinite impedance is that charge that is generated in the piezo-actuator due to external forces, is not removed from the actuator so it will add to the charge that was supplied by the amplifier. This results in a higher stiffness of the actuator in combination with the amplifier, as mentioned also in the previous section for an open connection. This increased stiffness gives a higher resonance frequency of the positioning system. The disadvantage of charge control, however, is that it is not DC-stable, showing more drift, and in general has more low frequency positioning noise than voltage control as the charge control involves integrated current-feedback operation of the amplifier.

In *voltage controlled* piezo-actuators the design of the power amplifier is

simpler, a better noise performance can be achieved, and the system shows less drift. The design of the power amplifier, which in this case has a low output impedance, and the better noise performance may give an advantage involving the system costs. A disadvantage of a voltage controlled piezo-actuator, however, is certainly the hysteresis which in this case may have to be compensated by other means.

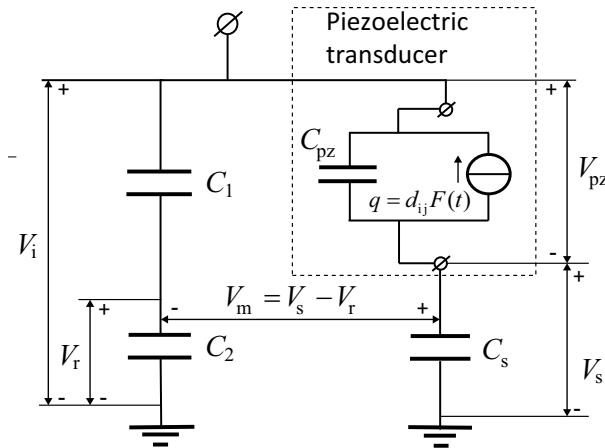
Another interesting design aspect in voltage controlled piezo-actuators is that the output impedance of the power amplifier forms together with the capacitance of the piezo-actuator a first-order low-pass filter. On the one hand this means that a potential bandwidth limitation may have to be considered in the system design, on the other hand this also gives a design freedom as adding a resistor in series with the piezo-element not only limits the maximum current, protecting the power amplifier and the actuator, but it also allows to shape the frequency response and introduce a roll-off before the resonance frequency of the actuator, which may be desirable particular in some open-loop positioning applications.

### 5.6.5.2 Self-sensing actuation

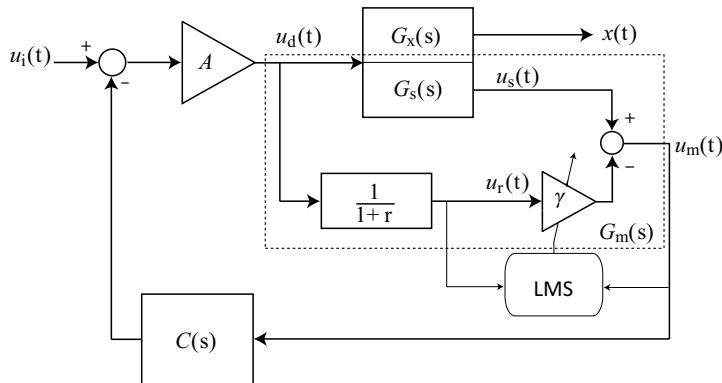
The transducer properties of piezoelectric material, as given by the constitutional laws and Equation 5.143, allows its simultaneous use as an actuator and a sensor.

Figure 5.50 shows the electric equivalent diagram of piezoelectric transducer, in this case the  $x$ -direction of the tube scanner depicted in Figure 5.47, which is integrated in a capacitive bridge circuit. The ratio of the capacitances in both vertical branches of the capacitive bridge is equal ( $C_1/C_2 = C_{pz}/C_s$ ). If the piezo-transducer is considered for a moment only as a capacitor  $C_{pz}$ , this implies that the charge distribution in the bridge circuit is in balance and no differential voltage  $V_m$  is measured across the bridge.

The piezo-transducer is in reality not a pure capacitor, because a force acting on the piezo-element will generate an additional charge  $q = d_{ij}F(t)$ , according to Equation 5.143. This changes the voltage across the piezo and leads to a charge imbalance in the bridge circuit, which can be observed by the measurement voltage  $V_m$  across the bridge. This means that external and reaction forces, as they occur at the load mass of the piezo, can be observed at the capacitive bridge circuit and eventually can be used for force feedback operation by feeding back the measurement voltage  $V_m$  to the power amplifier via a feedback controller, as shown in Figure 5.51. This feedback is based on the self-sensing capability of the piezo-transducer and allows for active damping of the resonances without the use of an explicit



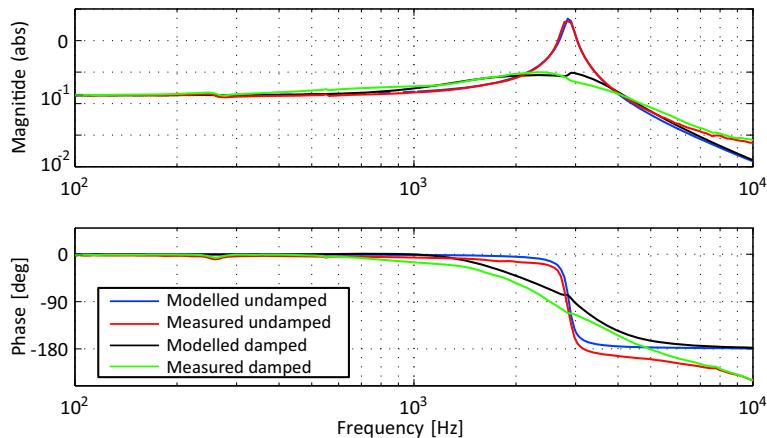
**Figure 5.50:** Electrical model of a piezoelectric actuator and capacitive bridge circuit to measure the charge that is generated by external forces on the actuator.



**Figure 5.51:** Schematic of the feedback control for active damping of the piezo-actuator via self-sensing actuation. The gain  $\gamma$  and the LMS-block form an adaptive balancing circuit to cancel gain variations due to the piezo-hysteresis, but have no direct effect on the active damping.

position sensor to measure the piezo-elongation, which is a very cost efficient implementation to improve the system performance.

A demonstration of active damping to reduce the resonance peak of the piezoelectric tube-actuator is shown in the Bode plot of Figure 5.52. The red solid line shows a measured frequency response  $G_x(s)$  of the piezo-tube without active damping. The blue line shows the frequency response of the



**Figure 5.52:** Bode plot of the measured and modelled frequency response of both the undamped and the actively damped piezoelectric tube-actuator.

second-order model that has been fitted to the measured data. Based on the fitted model a feedback controller  $C(s)$  has been designed to dampen the resonance peak of the tube scanner, resulting in the simulated frequency response of the controlled model as given by the black line in Figure 5.52. The green line of the measured frequency response of the transfer  $G_x(s)$  with active damping via the self-sensing actuation circuit confirms the reduction of the resonance peak by almost a factor ten, as predicted in the simulation.

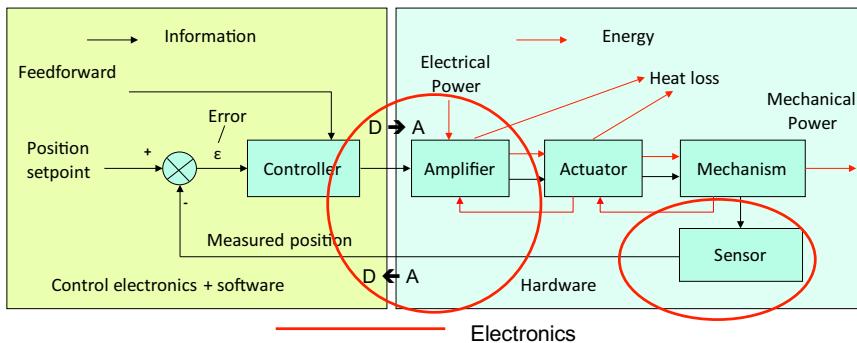
# Chapter 6

## Analogue electronics in mechatronic systems

Analogue electronic circuits are indispensable in the realisation of a mechatronic system. Both measurement and actuation need control of electric signals and knowledge about the important behavioural aspects of electronics is necessary to be able to design a balanced mechatronic system. This chapter will present low power electronics as applied in measuring, filtering and control and power electronics that are used for driving electromechanic actuators.

Generally electronics are divided in two different fields, digital and analogue. This distinction has developed over the years into two completely separated fields. Digital electronics consist of electronic switches that are designed to perform logical functions with binary numbers having only two states, “off” and “on”. These logical devices range from simple “AND” and “OR” operands to fully programmable digital signal processors and micro controllers. Because of this binary simplification it is hardly recognized anymore that all electronic switches show analogue behaviour. More so, it is the analogue behaviour, that determines the maximum frequency and switching speed of these devices. For this reason modern high speed electronic circuits like processors could only be designed with a clear knowledge of the analogue properties.

In Figure 6.1 it is shown, that next to power electronics also the control interface between the controller, the hardware and the observer part from the measurement sensor to the controller belong to the electronic domain in a mechatronic system. It will be demonstrated that even fast controllers are



**Figure 6.1:** Main locations of electronics in a mechatronic system.

best realised directly by analogue electronics. With the focus on these mainly analogue processes within a mechatronic system, analogue electronics determine the proportional behaviour of a dynamic system by amplification and filtering of signals.

Also another factor justifies the focus on analogue electronics in this book. The overwhelming amount of electronic products on the market, from serious systems to gadgets, work mostly with digital electronics. This abundance has resulted in a conviction with many people, that electronics are “ready” and that it is not necessary to investigate this technology anymore in order to be able to use it. Furthermore, the availability of computer simulation software has notably resulted in a reduction of the basic understanding of analogue electronics, because of lack of critical attitude. This has often resulted in mistakes when designing mechatronic systems. This observation is underlined by the fact, that even with the mentioned ample availability of electronics in our society, only very few people, outside the confined field of electronic expert, are able to combine the knowledge of electronics with the less abstract mechanical dynamics. It appears to be a real different domain for many mechatronic designers.

This large chapter consists of three sections. Section 6.1 on passive electronics concentrates on signal manipulation by filtering without amplification. The section includes impedances, sources, complex elements and passive filters. Also a parallel is drawn between the dynamics of an electronic filter and the dynamics of a mechanical system to show the similarity, in order to enlarge the mutual understanding of electronic and mechanical engineers. Section 6.2 consists of the important non-linear and active building blocks of an electronic circuit, followed by the design of low power amplifiers, both for small signal amplification and active filtering. Section 6.3 presents both

linear and switched-mode power amplifiers that serve as interface between the actuator and the controller. As is true for most of the chapters in this book, the entire field is large enough to fill many books without overlap, so necessary limitations are applied with a focus on basic understanding, rather than in depth analysis. Nevertheless the material as presented is sufficiently suitable for the basic design of electronics in mechatronic systems of average complexity.

## 6.1 Passive electronics

Analogue electronics consist of an active part and a passive part. The difference is, whether energy is added to the signal or not. An amplifier is a typical active component, while a resistor is a typical passive element. It is important to first understand passive circuits, before presenting the complexity of active electronics. It will show that even with only passive parts, several useful functions can be realised. This section starts with some considerations on electric networks.

### 6.1.1 Network theory and laws

Electric networks consist of a combination of electric sources and different passive components. The behaviour of the complete network is defined by several laws. This section describes these laws, starting with the abstract definition of the voltage and current source, as they are used as the signal input for the passive circuits that are presented next.

#### 6.1.1.1 Voltage source

In Chapter 2 the voltage source was introduced. An ideal voltage source has an internal electromotive force  $\mathcal{F}_e$  that is equal to the external voltage, even if a current is delivered to the load. As an example from real life, the mains power supply socket at home in most countries in Europe behaves almost like an ideal voltage source of approximately 230 V alternating current (AC) with a frequency of 50 Hz. This voltage remains almost constant even when more lighting or stronger appliances are switched on.

The word “almost” refers to the fact that “ideal” does not exist in the real world. Just imagine a voltage source of any arbitrary voltage that could deliver less than 1 mA or more than 1 GA without any effect on its voltage.



**Figure 6.2:** The different symbols of a voltage source, that are used in schematic diagrams of electronic circuits, show the sign of the voltage by means of a plus and minus. Sometimes the minus is left away. The most right version of these symbols is used in this book, as it clearly emphasises the zero source impedance of an ideal voltage source.

This would require the source to have the capability of delivering infinite power and that is unfortunately impossible. Nevertheless the example of the mains supply is a very close approximation of an ideal voltage source as long as the currents are kept within reasonable limits.

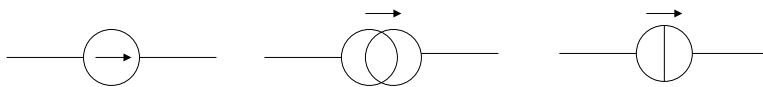
In a real voltage source, the external voltage is not fully equal to the electromotive force, due to an internal series impedance, originating from its physical properties. One can think of the internal wiring, the size of chemical electrodes and storage components and aspects related to the chemical separation of charge. This series impedance shows an internal electric field, opposite to the electromotive force, just like inside an external load. This field depends on the current and results in a change of the voltage difference over the electrodes of the source. In case of a resistive impedance, the voltage will be lower at an increased delivered current.

An ideal voltage source would have no internal series impedance and it can be approximated as a conducting wire when the influence of another electric source in the circuit has to be examined. This was shown in Chapter 5 about the interaction between an actuator and an amplifier, where the current from the two voltage sources could be derived separately by short-circuiting the other voltage source.

For this reason the version of the symbol that is located at the right in Figure 6.2 is used in this book when drawing a schematic diagram of an electronic circuit.

### 6.1.1.2 Current source

Most electronic circuits can be modelled with a voltage source, but sometimes it is more practical to work with the other extreme source of electricity, the current source. An ideal current source delivers a positive or negative current that is independent of the impedance of the conductive load and as a consequence the current is independent of the voltage. This functionality is less easy to understand from reasoning, when using the model of an internal



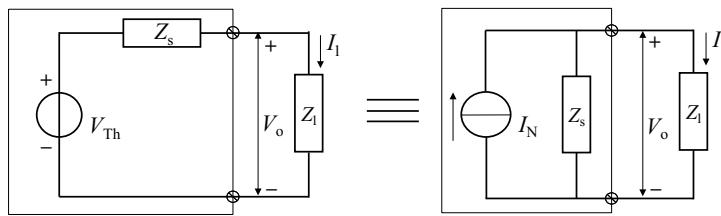
**Figure 6.3:** The different symbols of a current source show the direction of the current by means of an arrow. The most right version of these symbols is used in this book because of the clear semantics of the “blocking” line, representing the infinite internal impedance.

electrical field. Instead one might use the thinking model of a voltage source, that adapts its electromotive force actively to the load in order to keep the current at the same level. It will be shown that this indeed is a method to create a current source by active feedback control. When a current source is applied in a circuit, another external voltage source, like the movement induced voltage of an actuator, will not influence the current anymore. This implies that an ideal current source acts like an open wire for the current that is induced by other electric sources. In other words, a current source has an infinite output impedance. It is indeed true that the basic model of a current source is a voltage source with an infinite electromotive force and an infinite series resistance. This very artificial model is the reason for the preferred symbol as shown at the right side of Figure 6.3, as a current source can be replaced by an open connection in a circuit when the influence of another electric source has to be examined.

It is also obvious, that an ideal current source can not exist either. An ideal current source would imply for example that a continuous lightning flash over an infinite distance could be created. Nevertheless an almost ideal current source can be created that operates over a limited voltage range and it is a very important element in mechatronic systems in two ways. First of all, several sensors in measurement systems behave like a current source when the voltage is kept within a certain range. But even more important is the need for a current source when supplying an actuator with electrical power. In Chapter 5 it was shown, that a low impedance of the source results in damping and unwanted transmissibility of external vibrations through the actuator. With the application of a current-source power amplifier as will be presented in Section 6.3 that effect can be prevented.

### 6.1.1.3 Theorem of Norton and Thevenin

A real electric source can be modelled in two ways. The first consists of a voltage source in series with a finite impedance and is postulated by



**Figure 6.4:** Equivalent representation of real electric sources. The Thevenin model at the left consists of a voltage source with a series impedance. The Norton model at the right consists of a current source with a parallel impedance with the same value  $Z_s$  as with the Thevenin model. The relation between  $V_{Th}$  and  $I_N$  follows from Ohm's law:  $V_{Th} = I_N Z_s$ .

the French telegraph engineer Leon Charles Thevenin (1857-1926). The second method of a current source with that same impedance in parallel was postulated by the American electrotechnical engineer Edward Lawry Norton (1898-1983). Both versions are shown in Figure 6.4.

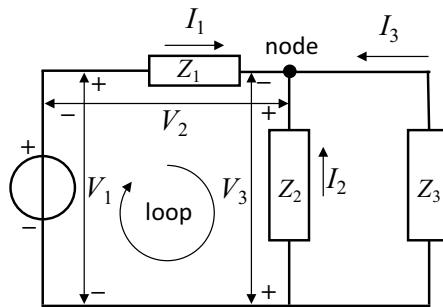
The Thevenin model is again the most straightforward to imagine. An external current  $I_l$  through a load  $Z_l$  will cause a voltage drop over the internal series impedance  $Z_s$ , resulting in a lower voltage over the electrodes of the source. The Norton model can be explained by defining the current  $I_N$  to be equal to the maximum current  $I_{l,max}$ , that can be delivered by the source in the Thevenin model. This maximum current occurs when the electrodes are short-circuited and equals with Ohm's law:

$$I_N = I_{l,max} = \frac{V_{Th}}{Z_s} \quad (6.1)$$

With this value, the voltage at the electrodes of both systems would be equal to  $V_{Th}$ , when no load is connected to the electrodes. As soon as a load  $Z_l$  starts to pull a current  $I_l$ , the voltage at the electrodes of the Thevenin model will drop with  $Z_s I_l$ , due to the series impedance  $Z_s$ . With the Norton model, the same current  $I_l$  will reduce the current through the parallel impedance  $Z_s$ , giving the same voltage drop over the electrodes.

#### 6.1.1.4 Kirchhoff's laws

Besides Ohm's law, as described in Chapter 2, two other important laws regarding electronic circuits need to be mentioned. They were named after Gustav Robert Kirchhoff (1824-1887), the German physicist who postulated them. Kirchhoff's laws are related to the currents and voltages in a network of electronic impedances. The first of Kirchhoff's laws deals with the currents



**Figure 6.5:** The laws of Kirchhoff. The currents at a node in a circuit add to zero ( $I_1 + I_2 + I_3 = 0$ ). Also the voltages over a closed-loop add to zero ( $V_1 + V_2 + V_3 = 0$ ). As long as the voltage and current directions are well defined, these laws help to calculate voltages and currents in complex electrical networks.

at a connecting point in a network and is based on the understanding that charge can not be stored at a connection point.

$$\sum I_{\text{node}} = 0 \quad (6.2)$$

The second of Kirchhoff's laws states that the sum of voltages over any closed-loop in an electronic circuit can only equal zero.

$$\sum V_{\text{loop}} = 0 \quad (6.3)$$

This is shown as an example in Figure 6.5.

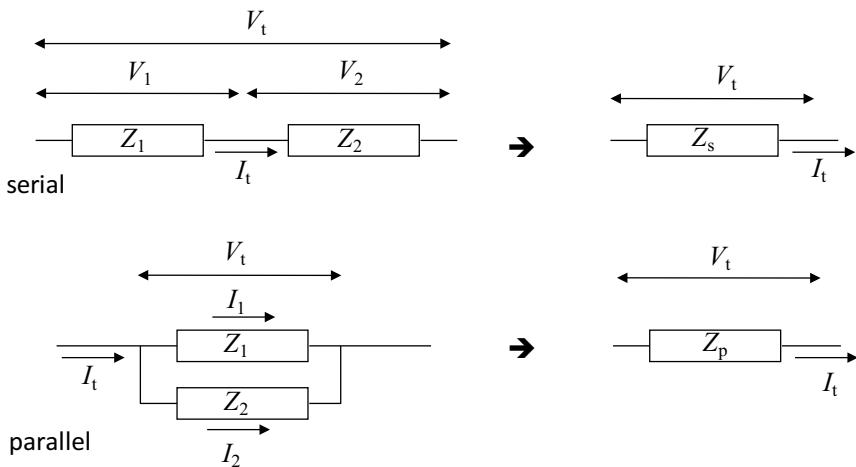
#### 6.1.1.5 Impedances in series or parallel

With the presented laws the currents and voltages in several electric circuits can be calculated. One example is the situation where different impedances are combined in series or parallel as shown in Figure 6.6. In the situation with two impedances in series, the current is shared and the impedance of the combination becomes:

$$Z_{\text{serial}} = \frac{V_t}{I_t} = \frac{V_1 + V_2}{I_t} = \frac{I_t Z_1 + I_t Z_2}{I_t} = Z_1 + Z_2 \quad (6.4)$$

With two parallel impedances, the voltage is shared and the impedance of the combination becomes:

$$Z_{\text{parallel}} = \frac{V_t}{I_t} = \frac{V_t}{I_1 + I_2} = \frac{V_t}{\frac{V_t}{Z_1} + \frac{V_t}{Z_2}} = \frac{1}{\frac{1}{Z_1} + \frac{1}{Z_2}} \quad (6.5)$$



**Figure 6.6:** Impedances in series simply add to the total impedance, as they share the same current, while the total voltage equals the sum of the voltage over each impedance. Parallel impedances share the same voltage, while the currents are divided over the impedances.

### 6.1.1.6 Voltage divider

Figure 6.7 shows a voltage divider that consists of two impedances. To calculate the output voltage, first the total current  $I$  from the source  $V_{\text{in}}$  into the series circuit  $Z_1 + Z_2$  is calculated and multiplied with  $Z_2$  to get the voltage over  $Z_2$ .

$$I = \frac{V_i}{Z_1 + Z_2} \quad V_o = IZ_2 = \frac{Z_2 V_i}{Z_1 + Z_2} \quad (6.6)$$

The relation between the output voltage  $V_o$  and the input voltage  $V_i$  equals:

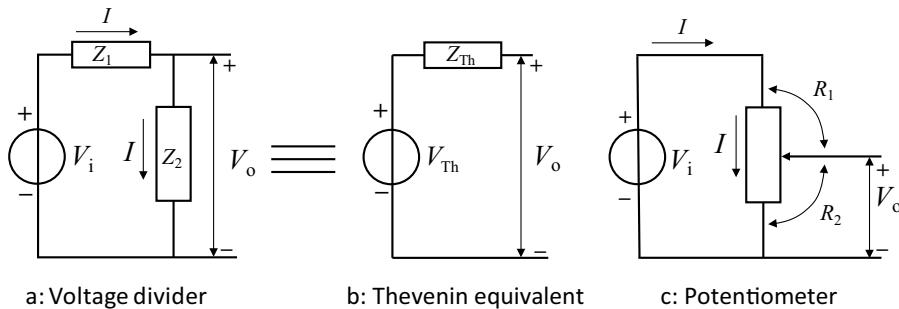
$$\frac{V_o}{V_i} = \frac{Z_2}{Z_1 + Z_2} \quad (6.7)$$

The equivalent circuit according to Thevenin is shown in the middle drawing of Figure 6.7.

The Thevenin voltage equals:

$$V_{\text{Th}} = V_{\text{in}} \frac{Z_2}{Z_1 + Z_2} \quad (6.8)$$

To find the Thevenin impedance a special insight is used, stating that this impedance is equal to the impedance of all paths inside the electric circuit

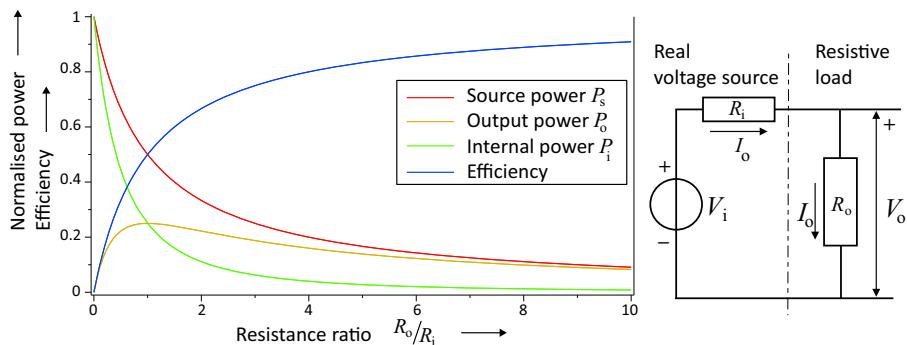


**Figure 6.7:** A voltage divider consists of two serial impedances and is based on the principle of the shared current. The voltage  $V_i$  gives a current  $I$  determined by the sum of both impedances, while the output voltage  $V_o$  is determined by the current and  $Z_2$  only. A special version of the voltage divider is the potentiometer, that can vary the ratio between its resistive impedances  $R_1$  and  $R_2$  to control the output voltage by a slider.

as observed from the outside. To determine that impedance a voltage source can be replaced by a conducting wire and a current source can be replaced by an open wire. In the situation of the voltage divider this means that  $Z_{Th}$  is equal to  $Z_1$  and  $Z_2$  in parallel as the impedance of the voltage source is zero:

$$Z_{Th} = \frac{1}{\frac{1}{Z_1} + \frac{1}{Z_2}} \quad (6.9)$$

With these values also the Norton equivalent of the voltage divider can be modelled according to Figure 6.4. A special execution of a voltage divider is the *potentiometer* as shown at the right drawing of Figure 6.7. A potentiometer consists of a resistor from the input to ground with a conductive slider that contacts the resistor on a location that is determined by an external rotating or translating element. The potentiometer creates a variable voltage at the output as function of the position of the slider.



**Figure 6.8:** The maximum power from a real voltage source is 25 % of the intrinsic power capability of the source and this maximum occurs when the external resistive load is equal to the resistive value of the source impedance. The efficiency is then only 50 %.

### 6.1.1.7 Maximum power of a real voltage source

A real voltage source consists of an ideal voltage source with an internal source impedance. It is interesting to investigate the maximum power that the source can deliver, because it is to be expected that the internal impedance will limit this power. In Chapter 2 it was explained that only resistive impedances dissipate power, which means that only the resistor value of the source impedance needs to be taken to calculate the internal power loss that is dissipated when a current is delivered to an electric load. This external load is for this simplified example approximated by a simple resistor and in that case the combined circuit behaves like a resistive voltage divider as shown at the right circuit diagram of Figure 6.8.

The output power in the load resistor equals the resistor value of the load times the current squared:

$$\begin{aligned} P_o &= I_o^2 R_o, \quad I_o = \frac{V_i}{R_o + R_i} \implies \\ P_o &= V_i^2 \frac{R_o}{(R_o + R_i)^2} = \frac{V_i^2}{R_o^2 + 2R_o R_i + R_i^2} = \frac{V_i^2}{R_o + 2R_i + \frac{R_i^2}{R_o}} \end{aligned} \quad (6.10)$$

This expression goes to zero for  $R_o = 0$ , which corresponds with the fact that then the output voltage is zero, and when  $R_o = \infty$  because then the output current is zero. The output shows a maximum when  $R_o = R_i$  which can be proven by finding the value of  $R_o$  that equals to a zero value of the

differentiation of the denominator of Equation (6.10) to  $R_o$ :

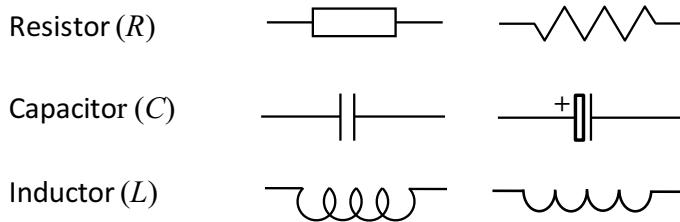
$$\frac{d \left( R_o + 2R_i + \frac{R_i^2}{R_o} \right)}{dR_o} = -\frac{R_i^2}{R_o^2} + 1 \quad (6.11)$$

which is zero for  $R_o = R_i$  corresponding with a maximum for the output power.

At the left side of the figure this useful output power is shown graphically together with the graph of the power delivered by the internal ideal voltage source ( $P_s = V_i I_o$ ), the power loss in the internal resistor ( $P_i = I_o^2 R_i$ ) and the efficiency being the output power divided by the source power. This graph shows that the source will deliver the maximum power with a very small value of the load resistance but most of this power is dissipated internally into heat. This is a phenomenon one can observe when short-circuiting a battery with sometimes hazardous effects when this is done with a Lithium based battery! It is also shown that the maximum useful output power equals only 25 % of the maximum power that the source can deliver.

The above reasoning should not be misunderstood. For a given real source the maximum power is obtained when the connected load impedance equals the internal impedance of the source. The efficiency is then however only 50 % and for a higher efficiency the load impedance needs to be significantly smaller than the source impedance. This means that from an efficiency point of view it is better to not use a power supply at its theoretical maximum power capability. With for example a battery it implies that the total energy should be drained over an extended period of time, which is mostly the case. When the battery is loaded at its maximum output power the output voltage would be half and then half the energy would be wasted in heat inside the battery.

An example where the maximum transfer of power is aimed for is in transport of signals over a long cable that is terminated with its *characteristic impedance*, a term that will be explained in more detail in Chapter 8. With a cable of for instance a characteristic impedance of  $50 \Omega$ , the source before the cable and the load after the cable should both have an impedance of the same  $50 \Omega$  and in that case the maximum signal power is transferred that the source could deliver. This maximum power level guarantees a maximum signal to noise ratio.



**Figure 6.9:** Symbols of the passive elements or “building blocks” in an electronic circuit. The two symbols for a resistor are equivalent. The left symbol of the capacitor is a normal capacitor, where the right symbol is a polarised electrolytic capacitor. This type needs a unidirectional voltage difference to keep the high capacitance and avoid current leakage. The two symbols of the inductor are also equivalent.

### 6.1.2 Impedances in electronic networks

The complex impedance  $Z$  can have aspects of three different properties, a frequency independent part, corresponding with a resistor, a part where the impedance decreases with the frequency, the capacitor and a part where the impedance increases with the frequency, the inductor. Figure 6.9 shows the symbols of these elements as used in an electronic circuit diagram. Mostly the elements are indicated as separate items and as a consequence it looks like they are physically separable. This is not always the case. In Chapter 5 the electrical equivalent of an electromagnetic actuator was presented, consisting of a voltage source, an inductor and a resistor. These elements all were a property of the same element, the coil, so by definition they are not separable. In electronic circuits mostly the elements are indeed treated as separate functional building blocks and for most less-critical applications this is quite sufficient.

#### 6.1.2.1 Resistors

The simplest passive element is the resistor ( $R$  in Ohm ( $\Omega$ )) with a frequency independent impedance, because of its non-complex, real character. The relation between voltage and current is according to Ohm’s law and as a consequence of the real impedance, the voltage and current have a proportional relation without time dependency so they are always in phase. For that reason, the power dissipated in a resistor equals the scalar product of the momentary voltage and the current value.

$$P_1 = IV \quad (6.12)$$

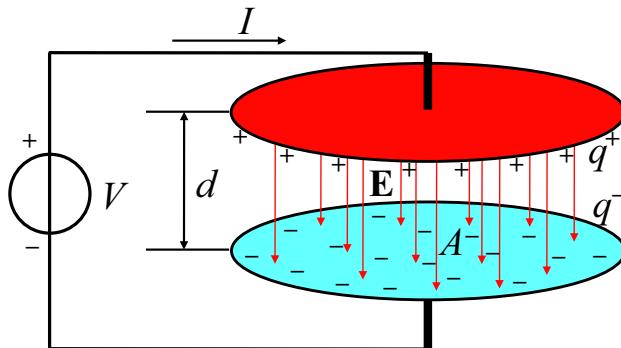


**Figure 6.10:** An overview of different resistors.

To get an idea about the variety of resistors, Figure 6.10 shows an overview of several types of resistors that are used in practical electronic circuits.

Resistors exist in an enormous range of resistance values from less than  $1\ \Omega$  to more than  $1\ G\Omega$ . This range is available in almost the same size, which makes them suitable in almost any application. Resistors are realised with different technologies, ranging from winding a metal wire with a high resistivity (wire-wound resistor) to the application of a thin carbon- or metal layer on a non-conductive ceramic round-or-rectangular body. In the latter case, the correct resistance value is obtained by trimming part of the resistive layer away with a laser. These carbon or metal film resistors are for instance recognisable in the figure by the colour code rings that indicate the value.

Over the years, the electronic industry has made their best effort to achieve almost ideal properties of all electronic elements. Nevertheless always some *parasitic* effect is present, that needs to be considered in critical applications. A resistor always has some capacitance and self-inductance, while requirements on these properties differ per situation. Next to the parasitic effects, the maximum dissipated power and the allowable temperature have given a boost to the enormous diversity in resistors. High power resistors are mainly wire wound, because of the insensitivity of the metal to high temperatures. These high power resistors are recognisable by their size, the often visible winding structure and the possible mounting to a heat sink, like shown with the green, the white and gold resistors on the photograph. As a



**Figure 6.11:** Basic principle of a capacitor.

consequence of the windings, wire wound resistors show a relatively high parasitic self-inductance, that makes them less suitable for high-frequency applications. With small high-frequency signals, the *Surface Mount Devices* (SMDs) are preferred as they can directly be mounted on a *printed circuit board* (PCB) without lead wires, that otherwise could introduce some self-inductance. SMD components are most frequently used in low power modern electronics and recognisable in the figure as the little white and blue blocks in the middle.

Although resistors have also some parasitic capacitance, depending on the size and the type, that capacitance is only important when really high frequencies are used like in precision measurement systems where often low-frequency signals are modulated with a high-frequency signal.

### 6.1.2.2 Capacitors

A capacitor is a complex impedance as its impedance is frequency dependent. In electric circuits it is characterised by the symbol  $C$  and its value has the unit Farad (F), named after Faraday. A capacitor is a *reactive impedance* because it can store energy in an internal electrical field between a pair of parallel plates, the electrodes, as shown in Figure 6.11.

In Chapter 2 on electricity, the relation between an electric field and its potential difference was shown. When two parallel plates with a different electric charge are positioned very close to each other, like in a capacitor, the field, corresponding with a charge difference between these plates, will be approximately homogeneous. This can be reasoned from the superposition of fields by drawing the field of every infinitesimal charge element and adding the field arrows as vectors. Field arrows of all charge elements that

point parallel to the plates will cancel each other out and as a result the remaining field lines all run orthogonal to the surface of the plates.

With a given electric field between the two electrodes, the potential difference increases with the distance of the electrodes. This implies, that with a given potential difference, the magnitude of the corresponding electric field is inversely proportional to the distance of the electrodes. This also implies, that the amount of charge difference between one electrode to the other is inversely proportional to the distance. In a capacitor, the charge difference is the result of the external voltage and it is not difficult to imagine that with a larger surface ( $A$ ) proportionally more charge needs to be transported between the electrodes to retain the same electric field over the entire surface.

As a last factor that determines the behaviour of a capacitor, the insulating material between the electrodes has its influence, due to the *polarisability* of the material. This polarisability is related to the property of certain materials, to create an electric dipole. By directing itself along the electric field, with the positive side directed towards the negative electrode, these dipoles counteract the electric field and as a consequence they increase the amount of charge displacement needed to sustain the electric field. This is in some way comparable with magnetism, where the Weiss domains of a material with a high relative permeability  $\mu_r$  direct to the magnetising field  $\mathbf{H}$ , reducing that field relative to the magnetic field  $\mathbf{B} = \mu_0\mu_r\mathbf{H}$ <sup>1</sup>. The electric property, that corresponds with  $\mu$  is called the *electric permittivity* ( $\epsilon$ ) with units As/Vm. In vacuum the value is called  $\epsilon_0$  and it has a direct relation with the magnetic permeability of vacuum ( $\mu_0$ ), that was presented in Chapter 5 and the speed of light  $c$  in vacuum:

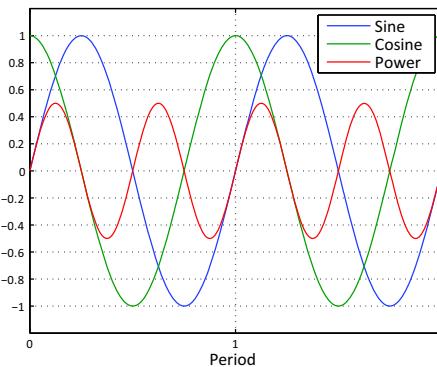
$$\epsilon_0\mu_0 = \frac{1}{c^2} \quad (6.13)$$

From this relation, when using the known values of  $c$  and  $\mu_0$  it follows that:

$$\epsilon_0 = \frac{1}{c^2\mu_0} \approx 8.8541878176 \cdot 10^{-12} \quad [\text{As/Vm}] \quad (6.14)$$

When the material between the plate is different from vacuum the total permittivity equals  $\epsilon_0\epsilon_r$  with the relative permittivity or *dielectric constant*  $\epsilon_r$  ranging from  $\approx 1$  for air to  $< 100$  for most solid materials and even higher for some ceramics and certain special polymers.

<sup>1</sup>Due to the different definitions this is somewhat confusing, as the Weiss domains help to increase the magnetic field  $\mathbf{B}$  relative to the magnetising field  $\mathbf{H}$ . It becomes comparable when the magnetic field is taken as the source of the magnetising field. As mentioned in Section 5.1.2, the cause and effect between both fields is a philosophical discussion as they simply co-exist.



**Figure 6.12:** The power of two multiplied sinusoidal functions with  $90^\circ$  phase difference averages to zero over time. The momentary power is alternating positive and negative. It represents the energy storage in the *reactive* impedance.

With this information it is possible to define the capacitance of a capacitor, which is its capability to store energy in an electric field by a charge displacement  $q_d$  at a given voltage:

$$C = \frac{q_d}{V} = \frac{\epsilon A}{d} \quad [\text{F}] \quad (6.15)$$

From a physical point of view it is important to note, that no charge is stored in a capacitor as a whole, but only at its electrodes. Charging a capacitor means storage of energy and displacement of charge. Due to the charge displacement a positive charge  $q_d^+$  is present at the positive electrode with an equal negative charge  $q_d^-$  at the negative electrode. The mentioned charge displacement  $q_d$  equals the integral of the current in the capacitor over time and because of this time relationship it is interesting to see how a capacitor behaves, when it is supplied with a changing (alternating, AC) voltage. The charge displacement is proportional to the voltage ( $q_d = CV$ ) and the current is equal to the rate of the change of charge so it is allowed to say:

$$I(t) = \frac{dq_d}{dt} = C \frac{dV(t)}{dt} \quad (6.16)$$

With this equation the voltage over a capacitor can be calculated at any time  $t_1$  when the starting voltage  $V_0$  at  $t_0$  and the current is given:

$$V(t) = V_0 + \frac{1}{C} \int_0^{t_1} I(t) dt \quad (6.17)$$

The energy stored in a capacitor is calculated by taking the integral of the power  $P_C$  needed to achieve the voltage over the time from  $t_0$ , where the voltage is zero, to  $t_1$ , where the voltage is  $V_1$ , and using Equation (6.16):

$$E_C = \int_{t_0}^{t_1} P_C(t) dt = \int_{t_0}^{t_1} V(t)I(t) dt = \int_{t_0}^{t_1} V(t)C \frac{dV}{dt} dt = C \int_0^{V_1} V(t) dV = \frac{1}{2}CV_1^2 \quad (6.18)$$

To determine the frequency domain behaviour of a capacitor, Equation (6.16) gives the complex impedance, by applying the Laplace transform and defining  $V_0 = 0$ :

$$I(t) = C \frac{dV(t)}{dt} \Rightarrow I(s) = C \cdot sV(s) \Rightarrow Z(s) = \frac{V(s)}{I(s)} = \frac{1}{sC} \Rightarrow Z(\omega) = \frac{1}{j\omega C} \quad (6.19)$$

This means that the impedance of a capacitor is a complex number, where the voltage shows a phase lag of  $90^\circ$  relative to the current and the impedance decreases proportional with increasing frequency, a +1 slope! When supplied from an AC source, the average power over time is zero because of this phase relationship, which follows from multiplying a sine and a cosine:

$$A_1 \sin(\omega t) A_2 \cos(\omega t) = 0.5 A_1 A_2 \sin(2\omega t) \quad (6.20)$$

This is illustrated in Figure 6.12 for  $A_1 = A_2 = 1$ . The power used to charge the capacitor alternates with an equal period with a negative power where the capacitor is discharged.

The capacitance value of a capacitor covers an extremely wide range from several Farad to smaller than one pico Farad. Initially large capacitors were very difficult to obtain. Just think of a metal plate of  $1 \text{ m}^2$  at a distance of  $100 \mu\text{m}$  in air. This results in a capacitance of only  $\approx 1 \cdot 10^{-7} \text{ F}$ . For this reason most practical capacitor values are given in either pF ( $10^{-12}$ ), nF ( $10^{-9}$ ) or  $\mu\text{F}$  ( $10^{-6}$ ). Figure 6.13 shows the large variety of types of capacitors. The largest capacitance values are achieved with electrolytic capacitors ranging from a few hundred nF to around  $10.000 \mu\text{F}$ , the big blue Philips capacitor, the green one and the tube with the nut. Electrolytic capacitors consist of two thin foils with a large surface of mostly aluminium, covered with an insulating oxide layer that acts as a high permittivity layer between the conductive foils. These foils are wound with a thin paper layer in between that is soaked with an electrolyte that is conductive and keeps the oxide layer in good order at the occasion of a damaged part in the oxide-layer. This self-healing effect improves their initially low reliability, but after some time the capacitance will inevitably reduce as a side effect of the renewed oxidation. The thin oxide layer with a high relative permeability is the cause



**Figure 6.13:** An overview of different capacitors.

for the initially high capacitance and for that reason electrolytic capacitors are mostly applied in power supplies to store energy.

Recent developments on a variation on electrolytic capacitors with an electrochemical double layer, have increased the maximum capacitance to over one Farad. Although their maximum voltage per cell is limited, the high energy density of these *super-capacitors* makes them applicable in high power applications as soon as all technological issues, regarding the combination of multiple capacitors for higher voltages, are solved.

Other smaller capacitors are made by winding thin foils of conductive material with intermediate non-conductive foils of a synthetic material with favourable dielectric properties, (the other round capacitors in the figure). Their capacitance ranges from around 100 pF to 100 µF. The smallest capacitor values are realised with single- or multilayer ceramic bodies with a metal layer as plate material. They can range between 1 pF to 1 µF.

Due to the winding and the thin foil, the electrolytic and foil capacitors both suffer from a parasitic resistance and some self-inductance. The parasitic resistance is represented by its equivalent series resistance (ESR), that limits the minimum impedance at high frequencies and induces a loss of power. For high-frequency applications, the ceramic capacitor is preferred as it shows the least parasitic inductance and resistance and it is also available in an SMD shape.

### 6.1.2.3 Inductors

The inductor consists of a coil with a certain self-inductance ( $L$  in Henry (H)) and was introduced in Chapter 5 as a property of electromagnetic actuators. An inductor is often named just shortly a coil or in some older literature a *solenoid*. In electronic circuits it acts as the second reactive element with a frequency dependent behaviour as it can store energy in a magnetic field.

As was presented in Chapter 5, the voltage over an inductor equals:

$$V(t) = n \frac{d\Phi_w}{dt} = \frac{d\Phi_{w,t}}{dt} \quad (6.21)$$

Where  $\Phi_w$  is the flux per winding and  $\Phi_{w,t}$  is the flux integrated over all windings ( $n\Phi$ ). The self-inductance was defined as the total amount of flux generated per Ampère:

$$L = \frac{\Phi_{w,t}}{I} = \frac{n\Phi_w}{I} = \frac{n^2}{R} \Rightarrow \Phi_{w,t} = LI \quad (6.22)$$

From this equation, combined with Equation (6.19), the Voltage as function of time becomes:

$$V(t) = \frac{d\Phi_{w,t}}{dt} = L \frac{dI(t)}{dt} \quad (6.23)$$

Also previously presented was the stored magnetic energy  $E_L$  in an inductor. Like with the capacitor, it can be calculated by integrating the power over the time from  $t_0$ , where the current is zero until  $t_1$ , where the current equals  $I_1$  for a certain value of  $L$ :

$$E_L = \int_{t_0}^{t_1} P_L(t) dt = \int_{t_0}^{t_1} I(t)L \frac{dI}{dt} dt = L \int_0^{I_1} I(t) dI = \frac{1}{2}LI_1^2 \quad (6.24)$$

In the frequency domain for AC currents the impedance of an inductor is obtained by the Laplace transform of Equation (6.23):

$$Z(s) = \frac{V(s)}{I(s)} = \frac{L \cdot sI(s)}{I(s)} = sL \Rightarrow Z(\omega) = j\omega L \quad (6.25)$$

This means that just like with the capacitor, also the impedance of an inductor is a complex number. In this case however, the phase of the voltage leads  $90^\circ$  to the current and the impedance increases with increasing frequency. Like with the capacitor, the average power over an inductor is zero, when supplied from an AC source, as was illustrated in Figure 6.12. In every period, the energy stored in the magnetic field in the inductor alternately



**Figure 6.14:** An overview of different inductors and ferromagnetic cores.

increases in the positive power cycle and decreases in the negative power cycle.

In reality, inductors are all made by winding wires around a core as can be seen in Figure 6.14. This core can consist of a material with a high relative permeability to achieve a high self-inductance or a low permeability for reason of linearity. For very high frequencies even a core can be omitted and only a section with a spiral track on a printed circuit board can suit the purpose. With some numbers in a real coil with  $n = 100$  windings, a core diameter ( $A$ ) of  $10 \times 10$  mm, a core length ( $l$ ) of 100 mm and a  $\mu_r$  of 100, the following relatively low value of the self-inductance is obtained:

$$L = \frac{n^2}{R} = \frac{n^2 \mu_0 \mu_r A}{l} = \frac{10^4 \cdot 4\pi \cdot 10^{-7} \cdot 10^{-4}}{0.1} = 1.25 \cdot 10^{-3} \text{ [H]} \quad (6.26)$$

For this reason the values of inductors are often given either in mH or  $\mu$ H.

Practical inductors suffer from many different parasitic properties. The intimately wound windings show a significant parasitic capacitance and resistance. When high currents are involved and a large self-inductance is needed, the size of the inductor will be large. Also the core gives many problems. Generally the relative permeability is depending on the momentary flux density by saturation. This means in practice, that the self-inductance is reduced with larger currents. And last but not least, the changing flux inside the core will induce eddy-currents within the core. This phenomenon was explained in Chapter 5 with the eddy-current ring in a Lorentz actuator. The eddy-currents in the core of an inductor will reduce the effective self-inductance and will result in dissipated energy, because the eddy-currents run in a resistive material. The eddy-currents can be limited by using a

non-conductive core material. An ideal group of materials in that respect are the ferrites, sintered material consisting of insulated, partly oxidised iron particles. Also a plastic compound of Iron particles with a thin oxide layer can be used. For low frequencies the eddy-current losses can be sufficiently limited by laminating the ferromagnetic core in thin sheets. Because the eddy-currents are running perpendicular to the flux lines the lamination should be in the direction of the flux to have the maximum effect.

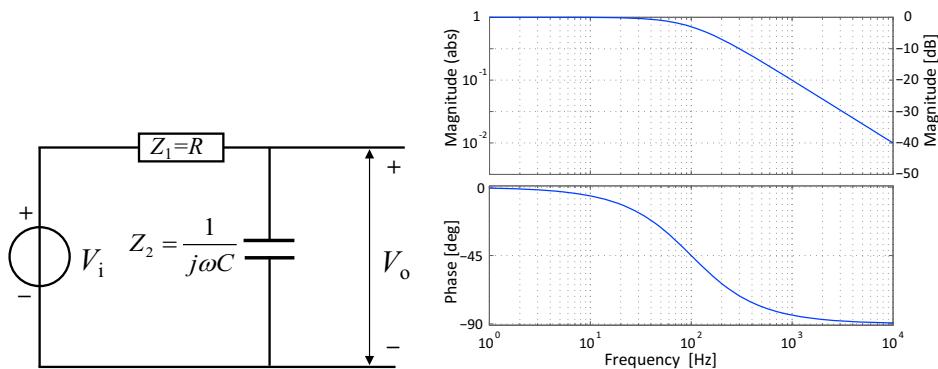
Another loss in the core is caused by the magnetic hysteresis. Contrary to permanent magnet materials, the ferromagnetic core of an inductor should ideally show no hysteresis. It is interesting to know that ferrites both can be designed to have a high hysteresis to create a permanent magnet, like mentioned in Chapter 5 as for having an extremely low hysteresis for inductor cores, all at acceptable cost levels. For laminated cores an alloy of iron with 2 – 5 % of silicon is a well-known material with a low hysteresis and is widely used in power transformers, rotating electrical machines and electromagnetic actuators.

### 6.1.3 Passive filters

Electronic filters are frequently applied in mechatronic systems for selecting frequency areas of interest, the *pass-band*, for rejecting frequency areas that introduce noise and other disturbances, the *attenuation-band* or for correcting phase properties in feedback systems. The field of electronic filters is very large and consists of passive filters with only resistors, capacitors and inductors and active filters with amplifiers. Passive filters are mainly used, when it is not possible to use an amplifier. This is for instance the case between an amplifier and an actuator like with the switched-mode power amplifiers of Section 6.3. But also with sensors at remote places at high temperatures passive filters can be useful.

#### 6.1.3.1 Passive first-order RC-filters

The most basic filter configuration is the *RC-filter* that is derived from the voltage divider of Figure 6.7, where one of the impedances is a frequency dependent component, like for instance a capacitor. The first filter to explain this principle is a first-order low-pass filter, that can be used for instance to reject noise above the *cut-off frequency* or *corner-frequency*  $\omega_0$  of the filter. This filter consists of a capacitor and a resistor as shown in Figure 6.15. As the impedance of the capacitor decreases with increasing frequency it can



**Figure 6.15:** An RC low-pass filter network and the Bode plot of its transfer function with  $R = 15 \text{ k}\Omega$  and  $C = 0.1 \mu\text{F}$ , giving a time constant  $\tau = 1.5 \text{ ms}$ . The increasing impedance  $Z_2$  of the capacitor reduces the magnitude of the output signal at frequencies above the corner-frequency  $f_0 = \omega_0/2\pi \approx 100 \text{ Hz}$ , where the magnitude of  $Z_2$  becomes larger than  $Z_1$ .

be reasoned, that this filter reduces the amplitude of the higher frequencies above the frequency, where the impedance of the capacitor becomes smaller than the impedance of the resistor. To derive the transfer function  $F(\omega)$  of this filter in the frequency domain Equation (6.7) of the voltage divider is used:

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2}{Z_1 + Z_2} \quad (6.7)$$

With  $Z_1 = R$ ,  $Z_2 = 1/j\omega C$ , the transfer function becomes:

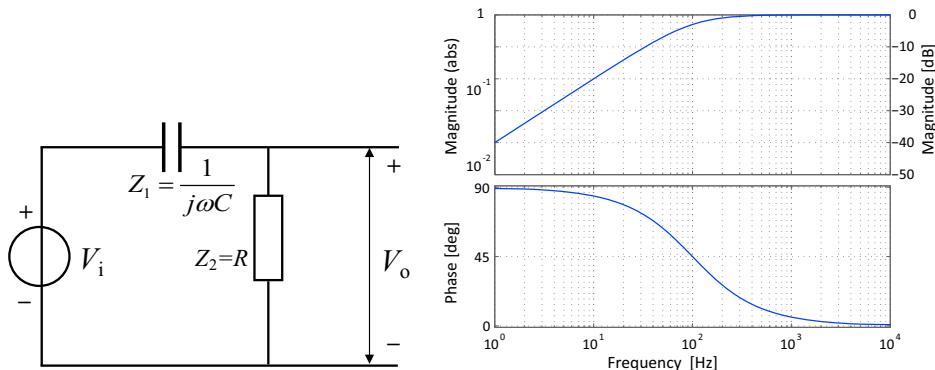
$$F(\omega) = \frac{V_o}{V_i} = \frac{\frac{1}{j\omega C}}{R + \frac{1}{j\omega C}} = \frac{1}{j\omega RC + 1} \quad (6.27)$$

At this point, the time constant  $\tau = RC$  also called the *RC-time* can be introduced. Because the transfer function of Equation (6.27) is dimensionless,  $RC$  is inversely proportional to the frequency  $\omega$ , so it has the unit of time. It is also known from control theory, that the time constant  $\tau$  defines the step response of a system starting at zero:

$$V = V_{st} \left( 1 - e^{-\frac{t}{\tau}} \right) \quad (6.28)$$

With the time constant  $\tau$ , Equation (6.27) becomes:

$$F(\omega) = \frac{1}{j\omega\tau + 1} \quad (6.29)$$



**Figure 6.16:** An RC high-pass filter network and the Bode plot of its transfer function with  $R = 15 \text{ k}\Omega$  and  $C = 0.1 \mu\text{F}$ , giving a time constant  $\tau = 1.5 \text{ ms}$ . The increasing impedance  $Z_2$  of the capacitor reduces the magnitude of the output signal at frequencies below the corner-frequency  $f_0 = \omega_0/2\pi \approx 100 \text{ Hz}$ , where the magnitude of  $Z_2$  becomes larger than  $Z_1$ .

To determine the amplitude at the corner-frequency  $\omega_0 = 1/\tau$ , the absolute value of the transfer function is calculated by taking the root of the sum of the squared real and imaginary term of the numerator and denominator:

$$|F(\omega_0)| = \sqrt{\frac{1^2}{1^2 + 1^2}} = \sqrt{\frac{1}{2}} = 0.707\dots \text{(in dB: -3 dB)} \quad (6.30)$$

The Bode plot of this filter is shown in Figure 6.15 for an example with  $R = 15 \text{ k}\Omega$  and  $C = 0.1 \mu\text{F}$ . This gives a time constant of  $\tau = 1.5 \text{ ms}$  and a corresponding corner-frequency of  $\omega_0 = 1/\tau \approx 660 \text{ rad/s}$  and  $f_0 = \omega_0/2\pi \approx 100 \text{ Hz}$ . The  $-1$  slope of the magnitude with  $-20 \text{ dB}$  per decade is as expected with a first-order system.

It is important to note that the characteristics of a passive filter are influenced by the load after the filter. When for example the filter of Figure 6.15 would be loaded with a resistive load, equal to  $Z_1$ , two effects would be observed. Firstly the gain of this filter in the pass-band would have a value of 0.5, because the load and  $Z_1$  would act like a frequency independent voltage divider. Secondly the cross-over frequency would be a factor 2 higher. Both effects can be calculated by first determining the equivalent Thevenin source voltage and impedance of  $Z_1$ , combined with the load and then combine that system with the capacitor.

To prevent this effect, these filters are commonly combined with buffer am-

plifiers with a very high input impedance and low output (source) impedance. These will be presented later.

The second filter of this kind is the high-pass filter, as used for instance to reject the DC value of a signal. Its basic design consists also on a capacitor and a resistor, but their location is exchanged as shown in Figure 6.16. It is clear, that at low frequencies the high impedance of the capacitor will prevent low-frequency current to flow into the output resistor.

Also with the high-pass filter Equation (6.7) is used to derive the transfer function  $F(\omega)$ :

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2}{Z_1 + Z_2} \quad (6.31)$$

With  $Z_1 = 1/j\omega C$ ,  $Z_2 = R$  and  $\tau = RC$  the transfer function becomes:

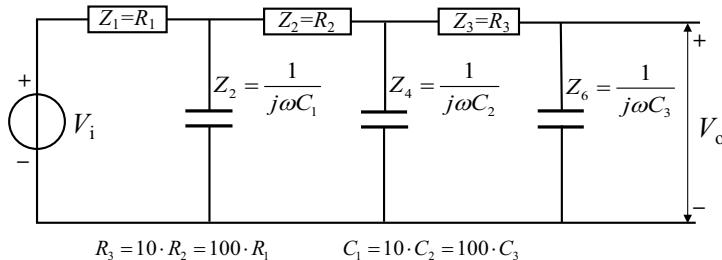
$$F(\omega) = \frac{V_o}{V_i} = \frac{R}{\frac{1}{j\omega C} + R} = \frac{j\omega RC}{j\omega RC + 1} = \frac{j\omega\tau}{j\omega\tau + 1} \quad (6.32)$$

The difference with the low-pass filter is the  $j\omega$  term in the numerator, corresponding with the differentiating Laplace variable  $s$ , which gives +1 slope with 90° phase lead and a 0 dB corner-frequency at  $\omega_0 = 1/\tau$ . This term, combined with the transfer function of the low-pass filter, results in the Bode plot of Figure 6.16 for the same values of the resistor and the capacitor.

Also with this filter a loading impedance over  $V_0$  will influence its characteristics. Even though the high-frequency pass-band gain will not be changed, the load impedance comes parallel to  $Z_2$  resulting in an increase of the cross-over frequency with more attenuation at lower frequencies.

### 6.1.3.2 Passive higher-order RC-filters

Often the attenuation of unwanted frequencies is required to be stronger than what is obtained with simple first-order filters. To create filters with a higher attenuation, a higher order filter needs to be applied. In principle this can be done by cascading several first-order filters. Figure 6.17 shows a third order configuration with a special measure to prevent mutual interference between the three filter sections. Normally the simple combination of identical passive filter sections introduces several problems, because each filter section would determine a load for the preceding section. To reduce these effects, the value for the resistor of each following segment can be chosen for instance a factor 10 higher than the resistor value in the preceding section, as shown in the figure. This helps reducing the problem, but it means that



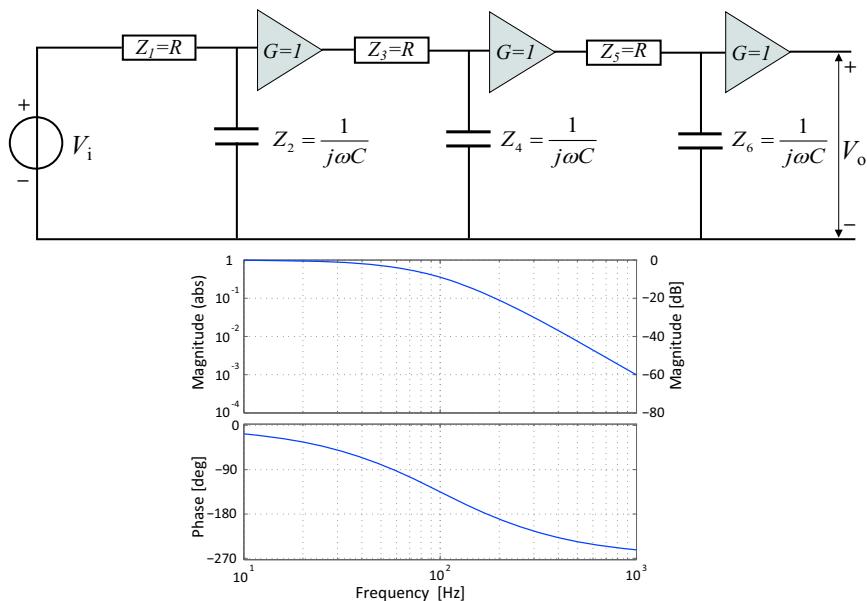
**Figure 6.17:** Creating a third order low-pass filter by cascading three first-order filters is possible with minimal mutual load. It is however better to introduce buffer amplifiers, as the output impedance of the filter becomes very high.

the effective output impedance of the combined filter would become very high. For that reason it is preferred to insert intermediate buffer amplifiers with a gain of one, like shown in Figure 6.18. How these amplifiers are made will be presented later in this chapter. For the explanation of the principle, it is sufficient to accept that these amplifiers enable to combine the separate sections without mutual influence, because they combine a very high input impedance with a very low output impedance. In that case the combined transfer function becomes a simple multiplication of the transfer functions of the separate filters:

$$F(s) = \frac{1}{(j\omega\tau_1 + 1)(j\omega\tau_2 + 1)(j\omega\tau_3 + 1)} \quad (6.33)$$

For an RC-time  $\tau_1 = \tau_2 = \tau_3 = 0.01$  s, the result is shown in the Bode plot of Figure 6.18. The figure clearly shows that the slope in the attenuation band is increased, but the transition to the pass-band is far from ideal. At the corner-frequency  $f_0 = \omega_0/2\pi \approx 100$  Hz, the attenuation is already 9 dB and the pass band bandwidth at -3 dB is at  $\approx 0.5 \times$  the corner-frequency, with a very gradual transition to a slope of -3.

As will be presented in the Section 6.2.5 on active filters, this trade-off between the properties in the attenuation-band and the pass-band can be optimised by the application of positive feedback, which is equal to increasing the imaginary term of the poles of the transfer function. When combining only first-order filters all poles are located on the real axis which implies a very high damping of the higher order filter. By relocating the poles, the gain at the cross-over frequency can be increased while keeping an acceptable steep slope in the attenuation band. It is worthwhile to add to this statement that it is not possible to create an ideal filter with an infinitely steep slope,



**Figure 6.18:** Cascading three first-order low-pass filters with intermediate buffer amplifiers results in a very gradual transition between the pass band below the corner-frequency and the attenuation band above  $f_0$ . At the corner-frequency the attenuation is already 9 dB. The Bode plot is drawn with  $R = 15 \text{ k}\Omega$  and  $C = 0.1 \mu\text{F}$ , corresponding with  $f_0 \approx 100 \text{ Hz}$ .

without phase and magnitude effects in the pass band and it will be shown that these electronic dynamics have a close relationship in behaviour with mechanical dynamics.

### 6.1.3.3 Passive LCR-filters

When it is not possible to use amplifiers, higher order filters can also be created by applying the other complex impedance, the inductor in an *LCR-filter*. When the resistor in the previous filters is replaced by an inductor, the filter effect will be enlarged, because the impedance of the inductor increases as function of the frequency.

#### Parallel resonant LCR-filters

In Figure 6.19 this second-order filtering effect is shown for a low-pass filter configuration, where the capacitor and inductor have a value of respectively 100  $\mu\text{F}$  and 25 mH.

The same approach as with the first-order filters is used to derive the transfer function:

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2}{Z_1 + Z_2} \quad (6.34)$$

With  $Z_1 = j\omega L$ ,  $Z_2 = 1/j\omega C$  the transfer function becomes:

$$F(\omega) = \frac{V_o}{V_i} = \frac{\frac{1}{j\omega C}}{j\omega L + \frac{1}{j\omega C}} = \frac{1}{\omega^2 LC + 1} \quad (6.35)$$

The corner-frequency of this filter is defined by:

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad (6.36)$$

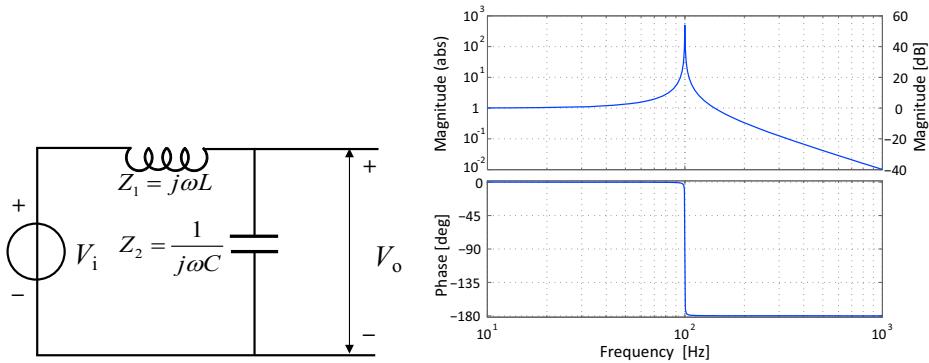
For the example of Figure 6.19 with a capacitor of  $100 \mu\text{F}$  and an inductor of  $25 \text{ mH}$ , the corner-frequency becomes  $\omega_0 = \sqrt{4 \cdot 10^5} = 630 \text{ rad/s}$  and  $f_0 = \omega_0/2\pi = 100 \text{ Hz}$ .

With the defined value of  $\omega_0$  the transfer function in the frequency domain is written as:

$$F(\omega) = \frac{1}{\frac{\omega^2}{\omega_0^2} + 1} \quad (6.37)$$

Similar to the mechanical mass-spring system in Chapter 3, this transfer function shows a resonance with infinite amplitude at  $\omega = \omega_0$ , while the corner-frequency of the filter is equal to the undamped natural frequency of the electronic resonator. In the mechanical situation, the height of this resonance is tuned by adding a damper. In electronic filters a resistor can act as a damping element as it dissipates energy. This resistive damping can be applied in two different ways. The first possibility is to apply the resistor at the output of the filter as a load. In that case the resistor will absorb more energy, when its value is low. The other possibility is to apply the resistor in series with the source, and in that case the resistor will take more energy when the resistance is high. This interesting contradictory effect will become more clear when the two configurations are modelled.

The first configuration is the previously shown LC low-pass filter with a damping load resistor at the output, as shown in Figure 6.20. This configuration is also called a *parallel-resonant* filter, because from the point of view of the impedances, the voltage source is a short-circuit. This means that all impedances are essentially connected in parallel.



**Figure 6.19:** A second-order LC low-pass filter network and the Bode plot of its transfer function with  $C = 100 \mu\text{F}$  and  $L = 25 \text{ mH}$ , giving a corner-frequency  $f_0 = 100 \text{ Hz}$ . Because of the lack of damping, the magnitude at the corner-frequency becomes infinite.

Again the same approach can be taken as with the previous filters to derive the transfer function, but in this case the impedance to ground consists of two parallel impedances  $Z_2$  and  $Z_3$ . The parallel configuration is noted with symbol ( $\parallel$ ):

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2 \parallel Z_3}{Z_1 + Z_2 \parallel Z_3} = \frac{1}{\frac{Z_1}{Z_2 \parallel Z_3} + 1} = \frac{1}{Z_1 \left( \frac{1}{Z_2} + \frac{1}{Z_3} \right) + 1} \quad (6.38)$$

With  $Z_1 = sL$ ,  $Z_2 = 1/sC$ ,  $Z_3 = R$ , the transfer function becomes:

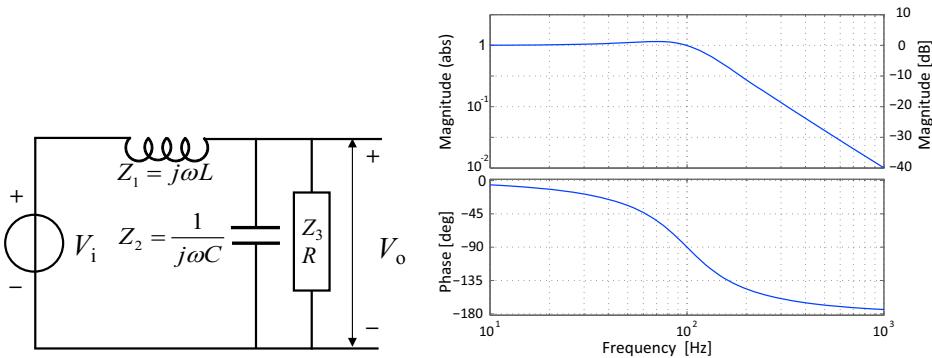
$$F(\omega) = \frac{V_o}{V_i} = \frac{1}{j\omega L \left( j\omega C + \frac{1}{R} \right) + 1} = \frac{1}{-\omega^2 LC + j\omega \frac{L}{R} + 1} \quad (6.39)$$

The same relation for the corner-frequency  $\omega_0 = \sqrt{1/LC}$  is used and the following relation for the quality factor  $Q$  and damping ratio  $\zeta$  is introduced:

$$Q = \frac{1}{2\zeta} = R \sqrt{\frac{C}{L}} \quad (6.40)$$

With these defined values the the transfer function in the frequency domain becomes:

$$F(\omega) = \frac{1}{-\frac{\omega^2}{\omega_0^2} + j \frac{2\zeta\omega}{\omega_0} + 1} = \frac{1}{-\frac{\omega^2}{\omega_0^2} + j \frac{\omega}{\omega_0 Q} + 1} \quad (6.41)$$



**Figure 6.20:** A second-order damped LC low-pass filter network and the Bode plot of its transfer function with  $100 \mu\text{F}$ ,  $L = 25 \text{ mH}$  and  $R = 16 \Omega$ . These values result in a corner-frequency  $f_0 = 100 \text{ Hz}$  and a damping with  $Q = 1$ ,  $\zeta = 0.5$ . A lower value of the resistor corresponds with an increased level of damping.

This represents a well controlled dynamic performance, when the damping ratio is chosen somewhere between 0.5 and 1. The resulting transfer function is flat in the pass-band and shows an almost constant slope in the attenuation-band. The Bode plot of this function has a clear resemblance with the compliance presentation in Chapter 3 of a mechanical damped mass-spring system. This mechanical versus electrical dynamic analogy will be shortly summarised a bit later in this chapter after completion of the presentation on the passive filters.

The parallel resonant configuration can also be used in a high-pass filter, by exchanging the location of the capacitor and the inductor, like shown in Figure 6.21.

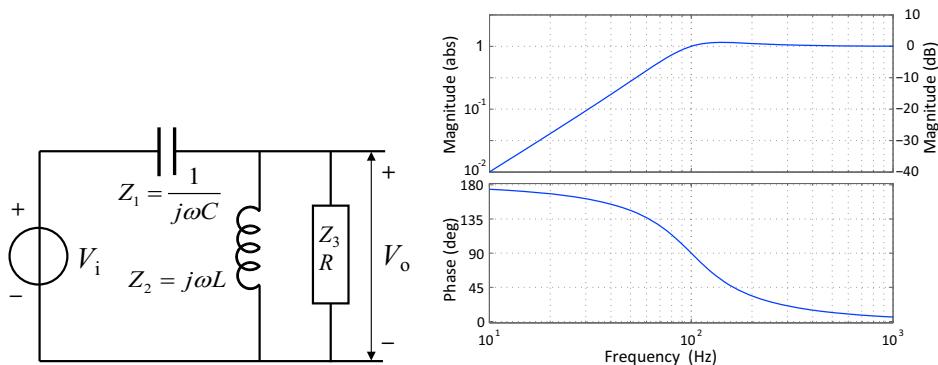
Again the same approach is applied to derive the transfer function:

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2 \parallel Z_3}{Z_1 + Z_2 \parallel Z_3} = \frac{1}{\frac{Z_1}{Z_2 \parallel Z_3} + 1} = \frac{1}{Z_1 \left( \frac{1}{Z_2} + \frac{1}{Z_3} \right) + 1} \quad (6.42)$$

With  $Z_1 = 1/sC$ ,  $Z_2 = sL$ ,  $Z_3 = R$ , the transfer function becomes:

$$F(\omega) = \frac{V_o}{V_i} = \frac{1}{\frac{1}{j\omega C} \left( \frac{1}{j\omega L} + \frac{1}{R} \right) + 1} = \frac{-\omega^2 LC}{-\omega^2 LC + j\omega \frac{L}{R} + 1} \quad (6.43)$$

Also the same relation for the corner-frequency  $\omega_0 = 1/\sqrt{LC}$  and damping ratio  $Q = 1/(2\zeta) = R\sqrt{C/L}$  is used and the transfer function in the frequency



**Figure 6.21:** A second-order damped LC high-pass filter network and the Bode plot of its transfer function with  $100 \mu\text{F}$ ,  $L = 25 \text{ mH}$  and  $R = 16 \Omega$ . These values result in a corner-frequency  $f_0 = 100 \text{ Hz}$  and a damping with  $Q = 1$ ,  $\zeta = 0.5$ . A lower value of the resistor corresponds with an increased level of damping.

domain becomes:

$$F(\omega) = \frac{-\frac{\omega^2}{\omega_0^2}}{-\frac{\omega^2}{\omega_0^2} + j\frac{2\zeta\omega}{\omega_0} + 1} = \frac{-\frac{\omega^2}{\omega_0^2}}{-\frac{\omega^2}{\omega_0^2} + j\frac{\omega}{\omega_0 Q} + 1} \quad (6.44)$$

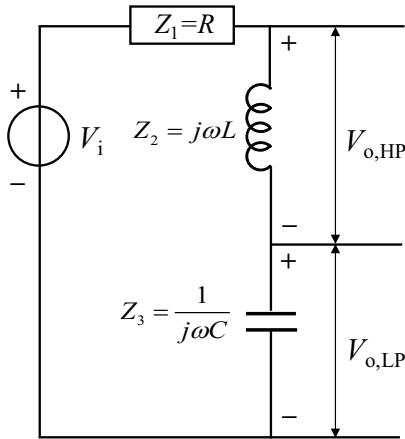
Like with the first-order filters, the only difference with the low-pass version is the numerator, where in this case the frequency terms are squared, corresponding with  $s^2$ . This results in a  $180^\circ$  phase shift (the minus term). The resulting Bode plot is in a dynamic sense completely comparable to the low-pass version with the same component values.

### Series-resonant LCR-filters

When a series resistor is applied to create the damping, the *series-resonant* circuit of Figure 6.22 is obtained.

The first observation in this configuration is that simultaneously a second-order low-pass and a high-pass filter<sup>2</sup> is obtained. All impedances are in series so they share the same current and the voltages add together to  $V_0$ . With these findings, the low-pass filter transfer function in the frequency

<sup>2</sup>The voltage over the resistor behaves like a first-order band-pass filter around the corner-frequency but it is left to the reader to examine this further.



**Figure 6.22:** A second-order damped LC low-pass and high-pass filter network in series configuration. A higher value of the resistor corresponds with an increased level of damping.

domain becomes:

$$F_{LP}(\omega) = \frac{V_{o,LP}}{V_i} = \frac{Z_3}{Z_1 + Z_2 + Z_3} = \frac{\frac{1}{j\omega C}}{R + \frac{1}{j\omega C} + j\omega L} = \frac{1}{-\omega^2 LC + j\omega RC + 1} \quad (6.45)$$

Likewise the transfer function of the high-pass filter function is:

$$F_{HP}(\omega) = \frac{V_{o,HP}}{V_i} = \frac{Z_2}{Z_1 + Z_2 + Z_3} = \frac{j\omega L}{R + \frac{1}{j\omega C} + j\omega L} = \frac{-\omega^2 LC}{-\omega^2 LC + j\omega RC + 1} \quad (6.46)$$

When comparing these equations with Equation (6.39) and Equation (6.43), the difference in the damping term  $RC$  versus  $L/R$  in the previous filters is clear. The quality factor for the series resonant filter can be determined by replacing the term in the parallel filter by the following steps.

The parallel quality factor equals

$$Q = R \sqrt{\frac{C}{L}} = \frac{R}{L} \sqrt{CL} \quad (6.47)$$

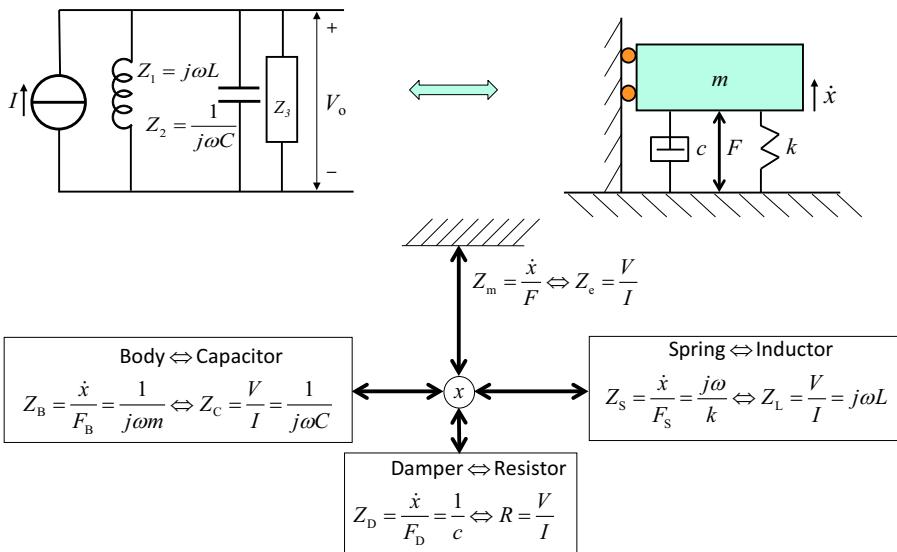
When exchanging  $L/R$  by  $RC$  the quality factor for the series resonant filter becomes:

$$Q = \frac{1}{RC} \sqrt{CL} = \frac{1}{R} \sqrt{\frac{L}{C}} \text{ with the damping ratio } \zeta = \frac{1}{2Q} = \frac{R}{2} \sqrt{\frac{C}{L}} \quad (6.48)$$

When these terms are applied, together with the unchanged  $\omega_0 = 1/\sqrt{LC}$  the same equations in the frequency domain are obtained as previously determined in Equation (6.41) and Equation (6.44).

### 6.1.4 Mechanical-electrical dynamic analogy

The previous presentation of passive second-order filters with an inductor, a capacitor and a resistor is so much similar to the presentation of the dynamics of a damped mass spring system that one wonders how these elements from electronics can be linked to their counterpart in the body, spring and damper of a mechanical dynamic system. Indeed this connection is possible and even more than one analogy can be defined of which one is presented here. In this equivalent dynamic system analogy, a parallel L-C-R-filter, driven by a periodically changing current source, is compared with the velocity response of a damped mass spring system driven by a periodically changing force, acting between the reference plane and the mass as shown in Figure 6.23. In this comparison first the current  $I$  is assumed to correspond with the force  $F$ . The logic of this assumption is based on the fact that at a connecting node of a mechanical system all forces add to zero, like in an electrical node all currents add to zero according to Kirchhoff's law. Also in a series configuration of mechanical elements the Force is shared, while in a series configuration of electric elements the current is shared. The next assumption is that the voltage  $V$  corresponds with the relative velocity  $v$ . This assumption is based on power as voltage times current and velocity times force both are equal to power. As a next step the impedances can be compared. Ohm's law in electricity can be replaced by a mechanical counterpart:  $\dot{x} = FZ_m$ , where  $Z_m$  equals the mechanical impedance. Like with electric impedances the three mechanical elements have their own typical impedance. The most straightforward is the damper with  $Z_D = 1/c$ , corresponding with the resistor  $Z_R = R$ . The frequency dependent impedance of a Capacitor  $Z_C = 1/j\omega C$  corresponds with the mechanical impedance of the body  $Z_B = 1/j\omega m$ . As the last of the elements, the spring corresponds with the inductor, not only because of their almost equivalent symbols but also by their behaviour. The mechanical impedance of the spring equals  $Z_S = j\omega/k$  corresponding with the impedance of the inductor,  $Z_I = j\omega L$ . These impedances show, that the capacitance ( $C$ ) corresponds with the mass ( $m$ ), the self-inductance ( $L$ ) with the compliance of the spring ( $C_s = 1/k$ ) and the Resistor ( $R$ ) with the inverse of the damping coefficient ( $c$ ) of the damper. To check this all in combination, the behaviour of the electrical circuit is first determined. According to Ohm's law the output voltage of the circuit



**Figure 6.23:** The analogy between a mechanical and electronic dynamic system, where voltage corresponds with velocity and current with force.

equals:

$$V_o = I(Z_1 \parallel Z_2 \parallel Z_3) \quad (6.49)$$

From this follows:

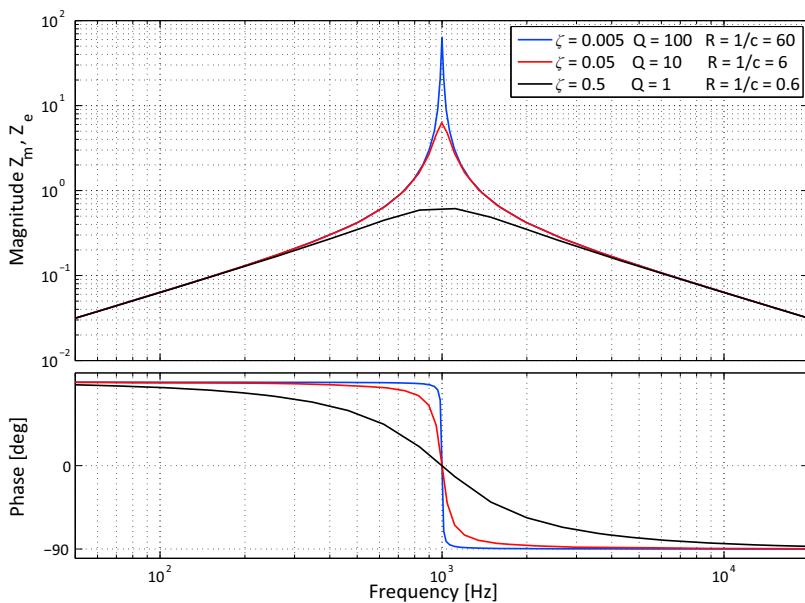
$$\frac{V_o}{I} = \frac{1}{\frac{1}{Z_1} + \frac{1}{Z_2} + \frac{1}{Z_3}} = \frac{1}{\frac{1}{j\omega L} + j\omega C + \frac{1}{R}} = \frac{j\omega L}{-\omega^2 LC + j\omega \frac{L}{R} + 1} \quad (6.50)$$

As a next step, the values for  $\omega_0$  and the damping ratio  $\zeta$  for both the electrical parallel resonant filter and the mechanical configuration are determined by mutually interchanging the mentioned terms:

$$\omega_0 = \frac{1}{\sqrt{LC}} = \sqrt{\frac{k}{m}} \quad \text{and} \quad \zeta = \frac{1}{2Q} = \frac{1}{2R} \sqrt{\frac{L}{C}} = \frac{c}{2\sqrt{km}} \quad (6.51)$$

With these terms, Equation (6.50) gives the following transfer functions for both the electrical and mechanical configuration:

$$\frac{V_o(\omega)}{I} = \frac{j\omega L}{-\frac{\omega^2}{\omega_0^2} + \frac{2j\omega\zeta}{\omega_0} + 1} \quad \text{and} \quad \frac{\dot{x}}{F}(\omega) = \frac{\frac{j\omega}{k}}{-\frac{\omega^2}{\omega_0^2} + \frac{2j\omega\zeta}{\omega_0} + 1} \quad (6.52)$$



**Figure 6.24:** Bode plot of the equivalent electrical impedance  $Z_e = V/I$  and the mechanical impedance  $Z_m = \dot{x}/F$  for  $1/k = L = 10^{-4}$  and  $m = C = 2.5 \cdot 10^{-4}$ . Below the corner-frequency of 1 kHz the line follows the spring/inductor line with a +1 slope and 90° phase lead. At the corner-frequency the impedance is real, which means that the level is only determined by the dissipative element, the resistor or the damper. Above 1 kHz the line follows a -1 slope, according to the mass/capacitance.

As an illustration, a Bode plot of the impedance of both systems is shown in Figure 6.24 with some values for the example. In the electrical domain the plot shows a resonance at the corner-frequency  $f_0 = \omega_0/2\pi = 1000$  Hz, with a resistive maximum depending on the resistor value. An application of such a circuit can be in a filter to reject a small frequency area, a band-reject filter, when the voltage divider of Figure 6.7 is used and  $Z_1$  is replaced by this circuit. When  $Z_2$  is replaced by the circuit a band-pass filter will result, that attenuates all frequencies outside of the corner-frequency of the circuit. The equivalent transfer function of the mechanical impedance corresponds with the velocity response of Equation (3.58) from in Chapter 3. The +1 slope with a 90° phase lead below the corner-frequency corresponds with the spring impedance in the mechanical system and the impedance of the inductor in the electronic circuit. The -1 slope with a 90° phase lag above the corner-frequency corresponds with the mass impedance in the mechanical system and the impedance of the capacitor in the electronic circuit.

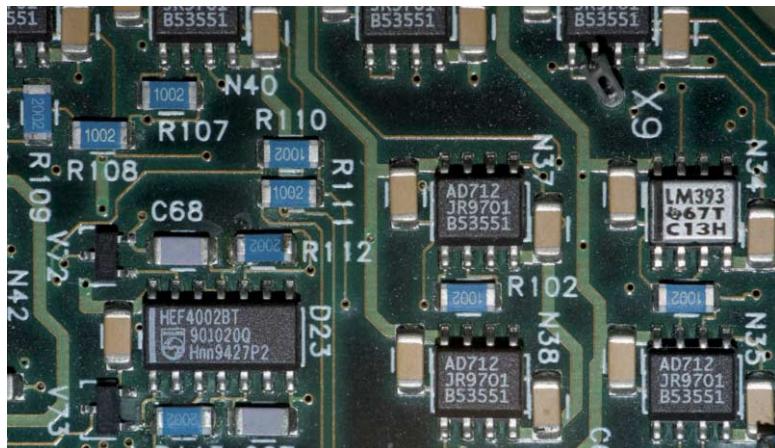
In Table 6.1 these findings are summarised.

**Table 6.1:** Electrical and mechanical analogy.

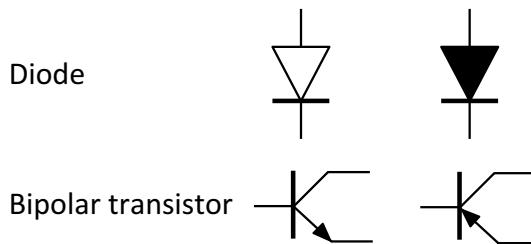
Mechanical	Electrical
Force ( $F$ )	Electric Current ( $I$ )
Speed ( $v$ )	Electric Voltage ( $V$ )
Mass ( $m$ )	Capacitance ( $C$ )
Compliance ( $1/k$ )	Self-inductance ( $L$ )
Damping coefficient ( $c$ )	Conductivity ( $1/R$ )
Energy ( $E = 0.5 \cdot mv^2$ )	Energy ( $E = 0.5 \cdot CV^2$ )
Power ( $P = Fv$ )	Power ( $P = IV$ )

## 6.2 Active electronics

Next to linear passive components, electronic circuits consist of different non-linear and active electronic building blocks. At the beginning of the age of electronics, the first of these active elements were based on the interaction between free electrons and charged metal elements in vacuum tubes. The discovery of the semiconductor diode effect in 1897 by the German physicist Karl Ferdinand Braun (1850 – 1918) started a cycle that would change the face of the world. The next significant event was the patenting in 1930 of a field effect transistor by the American physicist Julius Edgar Lilienfeld (1882 – 1963) who based his invention on a theoretical analysis without practical verification. A real working transistor was created more than 15 years later at Bell labs in 1947 by the American physicists William Bradford Shockley (1910 – 1989), John Bardeen (1908 – 1991) and Walter Houser Brattain (1902 – 1987) who won the Nobel prize for this important achievement. The real breakthrough in electronics came after the independent invention in 1958 of the integrated circuit by Jack Kilby (1923 – 2005), an American physicist, working at Texas Instruments, and by the American physicist Robert Noyce (1927 – 1990), co-founder of Fairchild Semiconductor. With these inventions the vacuum tubes became rapidly obsolete for professional applications. Presently the entire electronic field is divided in *discrete components*, with individual functional elements and the integrated circuits (ICs), that contain a multitude of different elements integrated on



**Figure 6.25:** Integrated circuits with passive components on a surface mounted printed circuit board.



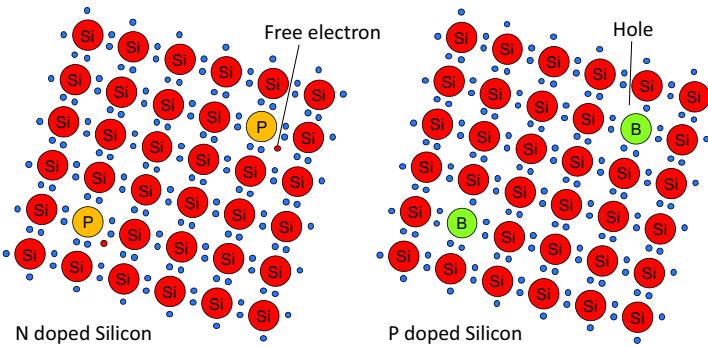
**Figure 6.26:** Two different symbols of a diode and a bipolar NPN (left) and PNP transistor. The arrows point towards the current conducting direction.

one single piece of Silicon, called a *chip*. These chips are packaged in mostly plastic housings with terminals, like shown in Figure 6.25. In the last few decades many versions of semiconductor-based discrete components have been created, ranging from the semiconductor diode and the bipolar transistor to special functional parts, like avalanche and zener diodes, the diac and triac, the field effect transistor (FET), the Metal Oxide Semiconductor FET (MOSFET) and manymore.

Integrated circuits are also diversified over many different applications, with a clear distinction between analogue and digital ICs. The best example of an analogue IC is the operational amplifier because of its applicability in a wide variation of electronic functions. Digital ICs are used for logical operations, digital processing and computer memory. In the past decades, their complexity has grown tremendously and presently they sometimes contain hundred million transistors or more, all connected by a huge network of integrated wiring. It would be impossible to give a complete overview of all elements and diversity of electronics in the context of this book. For that reason only the three most common examples of the discrete components are explained more in depth, the Silicon diode, which is a basic non-linear element, the Silicon bipolar transistor, that is used to create active amplifying electronics and the MOSFET that is used in digital electronics and in switched-mode power amplifiers.

### 6.2.1 Basic discrete semiconductors

The electronic circuit symbols for a diode and a bipolar transistor are shown in Figure 6.26. A diode allows an electric current to flow in one direction, but blocks it in the opposite direction. It can be seen as an electronic version of a check valve. Like a check valve that needs a pressure to open and shows a potential leakage in the blocking direction, also electric diodes do not



**Figure 6.27:** N- and P-material, created in a Silicon crystal structure, where one of the atoms is exchanged by Phosphor, to create N-material with one free electron or by Boron, to create P-material with one missing electron, a hole.

display a perfect on-off directionality but have a more complex non-linear electrical characteristic. In principle a diode is a passive semiconductor as it does not add energy.

The bipolar transistor is a real active component, that can be used to amplify signals by converting energy from a power supply into a signal with more power. Next to amplification a transistor can also be used as an electrically controlled switch. In this section first the working principle of a bipolar diode and transistor are explained based on the physics behind semiconductors in order to understand their main properties.

### Semiconductor physics

Semiconductors are chemically situated in the middle of the periodic system, which means that they can attain the ‘ideal’ state either by receiving  $n$  electrons, or by giving away the same amount of electrons. Examples of semiconductor materials are Silicon and Germanium. The most widely used material is Silicon, because of its excellent electronic properties, combined with its ample availability as the main constituent of sand.

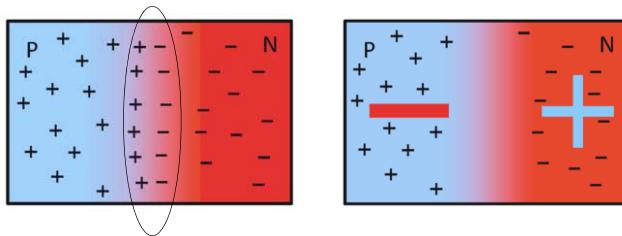
Pure Silicon has exactly 4 electrons per atom in its outer shell. Silicon would need to either receive 4 additional electrons or give these 4 electrons away in order to reach the “semi ideal” state. In the bulk material the 4 electrons of each atom are shared with their neighbours to achieve a true chemical bond. For this reason these electrons are firmly bound in the material structure. This immobility of the electrons means that pure Silicon behaves like an insulator. In order to achieve a semi-conducting state, it is necessary

to exchange some of the Silicon atoms by near neighbours in the periodic system, that either have three electrons in their outer shell, like Boron (B), or five electrons, like Phosphor (P). This exchange of atoms is called ‘doping’ and the resulting structure is shown in Figure 6.27.

Depending on the doping material, the doped Silicon shows a different behaviour. In case of doping with Phosphor atoms, additional not-bound, *free electrons* are made available. For that reason the silicon becomes an *N-material*. When the doping is done with Boron, so called *holes* are created, missing electron places. These electrons and holes are called *charge carriers*, as the doping in the material composition creates the possibility for the semiconductor to conduct current by means of these carriers. This conductivity is caused by the mobility of the free electrons and the holes. The mobility of the electrons is easy to understand, as they are not rigidly connected to an atom core, while the mobility of the holes is a bit strange and can be explained as follows. The presence of the hole disturbs the energy states in the material in such a way, that the normally bound electrons of the adjacent Silicon atoms tend to jump onto that hole, when an internal electric field drives it in that direction. As a result this electron leaves an empty place, a new hole, where the electron came from. This new hole can be seen as a displacement of the original hole in the opposite direction of the “jumped” electron. For this reason, charge transport by holes takes place in the same direction as the current. Note that due to this mechanism, also the charge transport by holes is in fact achieved by electrons. The difference is, that the charge transport by the holes in the P-material is done by normally bound electrons from the Silicon, while the charge transport by the electrons in the N-material is done by the free electrons, coming from the doping atoms.

Before entering the next section, where N- and P-material will be combined to create a diode, it is important to add some additional remarks:

- Although the P-material has free holes and the N-material free electrons, those materials are not charged as they originate from a not charged doping atom added to the Silicon crystal. This is fully identical to an uncharged metal that contains many free electrons.
- The concentration of the doping elements is very low. This concentration varies between approximately  $10^{-6}$  and  $10^{-8}$  depending on the purpose.
- The speed of the conduction in P-material is lower than the speed of conduction in N-material, due to the indirect conduction mode of holes. This difference is observed in high-frequency applications.



**Figure 6.28:** Combining N- and P-material creates a depletion layer, because electrons of the N-material recombine with the holes of the P-material. As a result, the N-material becomes positively charged and the P-material becomes negatively charged, according to the large plus and minus sign. This results in a measurable potential difference between both materials. The shown small plus and minus signs represent the remaining free moving charges.

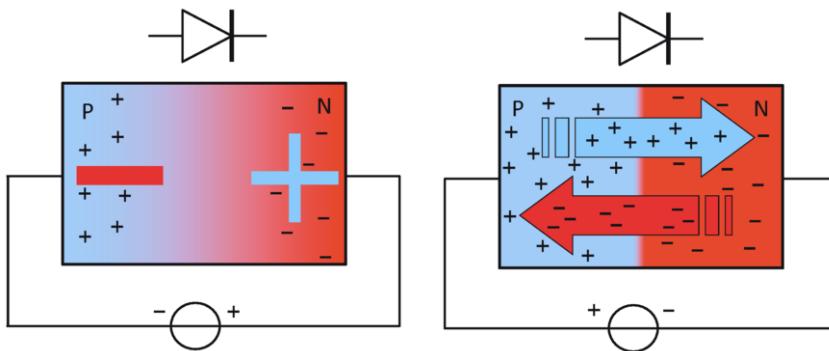
### 6.2.1.1 Semiconductor diode

When N- and P-material are brought into electrical contact, the electrons of the N-material and the holes of the P-material will diffuse into each other and the electrons will fill the holes at the interface area. As a result an intermediate layer is created without free electrons and holes. This process is called the *recombination* of the electrons and the holes and this layer is called the *depletion layer*, as shown in Figure 6.28. This layer behaves like an insulator due to the lack of free movable charge carriers.

Because the P-material loses positively charged holes and the N-material loses negatively charged electrons, the P-material becomes negatively charged and the N-material becomes positively charged, creating an electromotive force, that can be detected by its external potential difference.

In Figure 6.29 it is shown what happens, if an external voltage source is applied between the combined N- and P-material. At the left the external voltage is in the same direction as the electromotive force, caused by the recombination of the charge carriers. The external source adds electrons to the P-material and takes electrons from the N-material. This results in even less free charge carriers, and because of this enlarged depletion layer no current can continue to flow. The thickness of the depletion layer becomes proportional to the external voltage and the internal electromotive force that corresponds to this large depletion layer is in equilibrium with the externally applied voltage.

At the situation of the right of Figure 6.29, the external voltage is applied in a direction opposite to the electromotive force of the recombined depletion

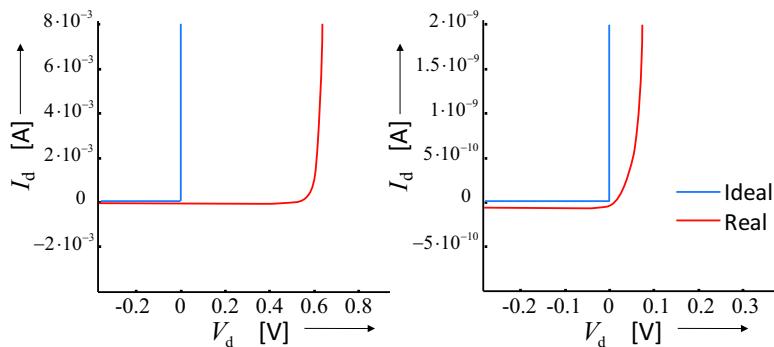


**Figure 6.29:** A semiconductor diode conducts current in one direction, when the external voltage is directed opposite to the internal electromotive force and blocks it in the other direction. In the conductive mode both the electrons and the holes contribute to the current.

layer. In that situation electrons are added to the N-material and taken from the P-material, compensating the effect of the recombination. As a result the depletion layer disappears and a conductive path is created where both the holes and the free electrons contribute to the current.

It is obvious that these effects don't happen instantaneous without any transition phenomena between conducting and blocking. This is illustrated in Figure 6.30, that shows the real characteristics of a typical Silicon semiconductor diode. In the conductive direction it is to be expected, that the internal electromotive force of the depletion layer determines a certain threshold level, that the external voltage has to surpass in order to make the diode conducting. Furthermore, due to non-ideal material properties, always some leakage current is present in the blocking direction. This leakage current also reduces the external voltage that can be measured at the terminals of a not connected diode below the electromotive force caused by the depletion layer. For a silicon based diode the threshold voltage  $V_{th}$  is around 0.6 – 0.7 V in practical situations. This value is strongly temperature dependent, because it is determined by the mobility of the charge carriers, that cause the depletion layer. This mobility is increased at elevated temperatures.

Another side effect is caused by the fact, that the depletion layer is a non-charged (insulating) layer between two parallel plates, which acts like a parasitic capacitor. The need to charge this capacitor determines the current needed to change the diode from the conductive state into the blocking



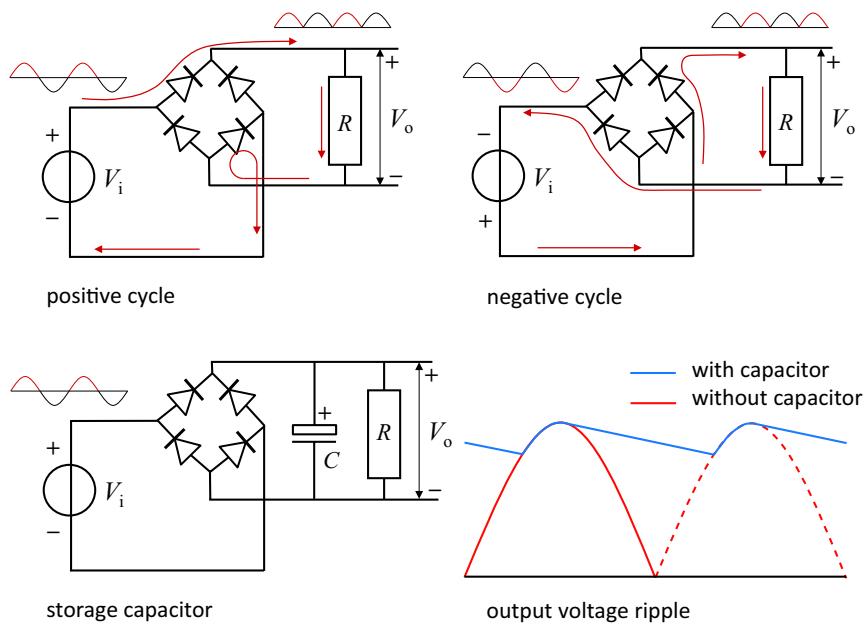
**Figure 6.30:** Real versus ideal characteristics of a silicon diode. A threshold voltage  $V_{th}$  of approximately 0.6 V is present in the conductive direction at moderate current levels above the mA range, as shown at the left. At extremely low current levels in the nA range also some leakage current occurs with a voltage in the reverse direction.

state and vice versa, when the current is reversed. This is an important characteristic, because diodes are frequently used to rectify alternating currents into direct currents in power supplies and high-frequency signal demodulators.

Figure 6.31 shows an example of the most frequently used rectifier circuit, the *bridge-rectifier*, that consists of 4 diodes and a storage capacitor. As can be seen in the drawing both the positive and the negative voltage from the input results in a unidirectional positive voltage at the output. The large storage capacitor will flatten the pulsating voltage into a DC voltage with a ripple that depends on the current in the load and the capacitance value of the storage capacitor. This can be filtered further with a passive low-pass filter, as was presented in the previous part.

In the previous chapter the electric transformer was presented. One of the conclusions was that a high-frequency AC voltage is necessary to be able to reduce the size. The rectification of a high-frequency AC voltage requires very fast switching diodes in order to reduce the power dissipation, related to the continuous charging and discharging of the intrinsic parasitic capacitor from the depletion layer. This dissipation is caused by the unavoidable parasitic source resistance in the circuit, that causes the power loss. Another application of fast diodes is in the switched-mode and resonant amplifiers that will be presented in the section on power amplifiers.

Apparently even this single component can be optimised for many different applications.



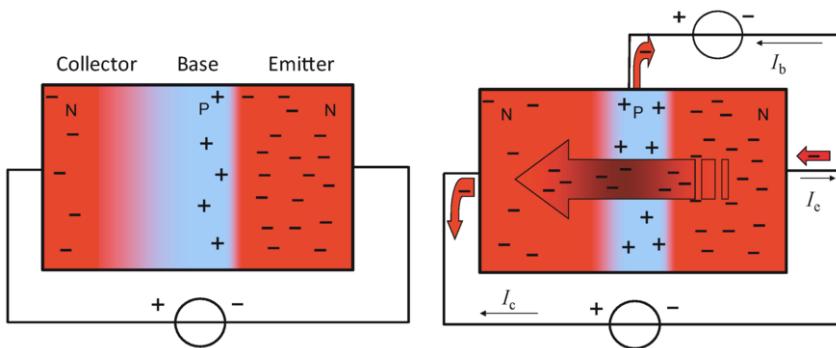
**Figure 6.31:** Diode bridge rectifier for AC voltage. Both the flow of current at the positive and the negative cycle of the input voltage result in a unidirectional current in the load, because of the diode configuration. By adding a storage capacitor a DC voltage is obtained with only a limited AC ripple. The capacitor is charged in the positive slope of the rectified input signal and discharged by the load in the negative slope.

### 6.2.1.2 Bipolar transistors

A bipolar transistor is a combination of three layers of successive N- and P-material. Depending on the order of the layers with N- and P-material between the emitter and the collector, the transistor is an NPN or a PNP type.

To explain the working principle, a simplified thinking model of an NPN transistor is shown in Figure 6.32. When a positive voltage is applied between the collector and the emitter, the interface between the collector and the base will be non-conductive and the interface between the base and the emitter is in the conductive mode. The net result is that the transistor is not conducting as the layers are connected in series.

As soon as a positive current is applied between the base and the emitter, electrons start to move from the emitter to the base in the conductive direction of the base-emitter PN-junction. The base is extremely thin and the



**Figure 6.32:** Combining N-, P- and N-material gives an NPN transistor. Without base current  $I_b$  the transistor is not conducting. A base current in the conductive direction of the base-emitter junction causes a large number of electrons to pass the collector-base barrier, resulting in a collector current that is larger than the base current.

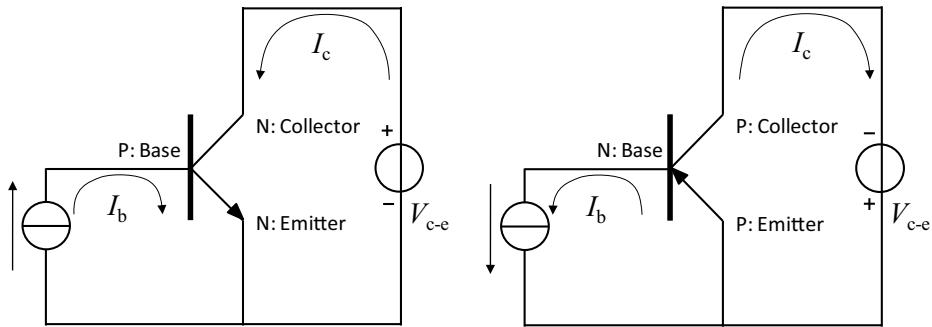
high voltage at the collector determines a strong electric field, accelerating the electrons that move from the emitter towards the base to fill the holes that were created by the base current. Because of the accelerated state of these *hot electrons*, not all of them recombine with holes in the base, but the majority breaks through the barrier between the base and the collector. In this way an electron flow is generated from emitter to collector, equivalent to a current flow from collector to emitter. This current is larger than the current in the base that caused it. Increasing the current on the base ( $I_b$ ) will trigger more emitter electrons, that break through the collector-base junction.

With this principle it is explained that a transistor acts as a current amplifier, with a current-amplification ratio  $\beta$ , also called  $h_{fe}$ , between the collector and base current:

$$\beta = h_{fe} = \frac{I_c}{I_b} \quad (6.53)$$

The example was given for an NPN transistor, where the electrons play the main role. For that reason in that configuration the electrons are called the *majority carriers* and the holes are the *minority carriers* as they only play a role in the base current.

With a PNP transistor it is just the other way around. The holes are the majority carriers and all currents run in the opposite direction of the NPN



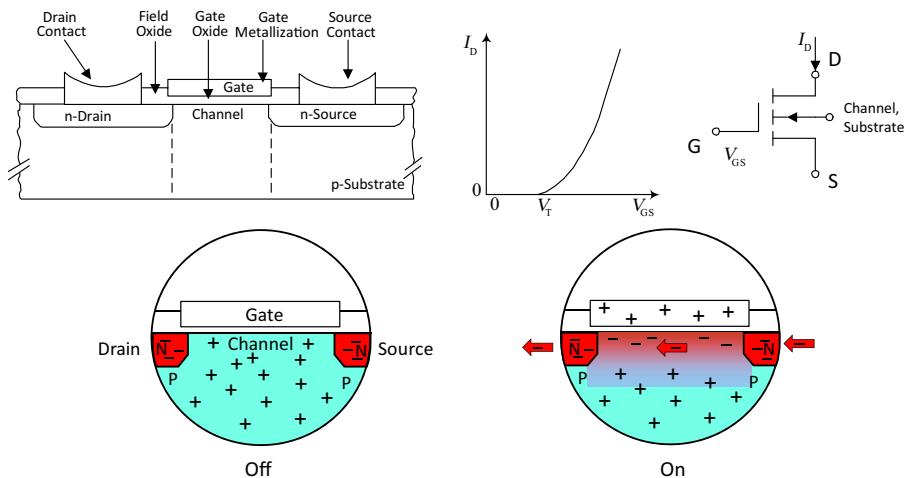
**Figure 6.33:** Current directions in an NPN and PNP transistor.

example as shown in Figure 6.33. Because of the different character of the majority carriers, free electrons or holes, both types of transistors are never fully identical, aside of the current direction. It is easier to achieve a high  $\beta$  with an NPN than with a PNP transistor and holes are slower than electrons. Still this choice enables the electronic designer to apply these different transistors in almost symmetrical circuits that control both positive and negative currents.

### 6.2.1.3 MOSFET

The name MOSFET means Metal Oxide Semiconductor Field Effect Transistor. It is a transistor that operates because of a externally controllable electric field, while it uses a metal oxide to insulate the input where the controlling signal is applied. The electrodes of a MOSFET have different names than with a bipolar transistor. The input electrode is called the *Gate* and is equivalent in function to the base in a bipolar transistor. The *Drain* is equivalent in function to the collector as it drains the electrons in an N-channel MOSFET, which is the most common version and the *Source* is equivalent in function to the emitter as it is the source of the electrons in an N-channel MOSFET. In order to distinguish it from a voltage or current source, the *Source*, *Gate* and *Drain* of a MOSFET will be written as a name starting with a capital.

MOSFETs are the most common semiconductors in integrated logic circuits, like microprocessors and memory, because of their speed, size and possibility to work with very low power supply voltage levels. The speed is high because the current flow takes place by majority carriers only, which are electrons in N-channel MOSFETs and as mentioned, charge transport by electrons is



**Figure 6.34:** A MOSFET operates by an electric field. With an N-channel MOSFET the channel is located in the substrate, made of P-type Silicon, between the Drain D and the Source S, both made of N-type Silicon. In an enhancement-mode MOSFET the channel is normally not conducting, but when a positive voltage is applied between the insulated Gate and the Source the corresponding electric field will drive away the holes from the P-type channel and attract electrons from the Source, that as a result will flow towards the Drain. The relation between the Gate to Source voltage and the Drain current is less steep than with a bipolar transistor and the behaviour is more like a variable resistor.  
(Courtesy of International Rectifier)

significantly faster than the indirect charge transport by holes. Next to the small signal type MOSFETs, also high power versions have been developed, that consist of millions of separate small MOSFETs connected in parallel. They are applied in high power switching applications and will be further presented in Section 6.3.

The structure of one type of N-channel MOSFET, the horizontal *enhancement-mode* type is shown in Figure 6.34. Many different types of MOSFETs exist, the *depletion-mode* MOSFETs are normally conducting, but the normally non-conducting enhancement-mode version is most frequently used. The horizontal type is the most easy to use for explaining the working principle, while the vertical type is used to create power MOSFETs.

The substrate of a horizontal N-channel enhancement-mode MOSFET consists of P-doped Silicon. The Drain and the Source are made of N-doped Silicon, separated by the substrate. The area in the substrate between the Drain and the Source is called the *channel*. Unlike with the bipolar

transistor, the depletion layer between the P-and N-material plays no other role than to insulate the Drain from the substrate when the voltage on the Drain is positive relative to the substrate. Normally the junction between the Source and the substrate is in the conductive direction during operation and for that reason they are mostly internally connected.

The real functionality takes place in the channel area around the Gate. The Gate is made of metal and is electrically insulated from the substrate, the channel and the Source by a metal-oxide layer. This gave the name "MOS" to the device, although often the metal of the Gate is replaced by high doped poly-crystalline Silicon and the oxide by Silicon oxide, because of the better compatibility with the Silicon manufacturing process. This makes however no difference for the working principle.

The oxide layer between the Gate and the channel is very thin and when a positive voltage is applied to the Gate, relative to the Source, the holes in the channel are pushed away due to the electric field by the Gate voltage, while the electrons are attracted. As a result, a large number of free electrons become available in the channel. Effectively the original P-doped substrate becomes a virtual N-material, with equal properties as the material of the Drain and the Source. This way, the channel becomes a conductive path between the Drain and the Source, with a width that is determined by the voltage level on the Gate. The resistance of the channel depends on the width and with this principle a MOSFET behaves like a controllable resistor. The depletion mode MOSFET works in the opposite way. In that case the channel is in principle conducting by a different doping, and with N-material of the channel a negative voltage on the Gate will repel the electrons, creating a depleted area with an increased resistance. Because this type is not used in amplifiers for mechatronic systems it will not further be treated.

In spite of the quite simple functionality of MOSFETs, bipolar transistors are still more often used in analogue amplifiers and for that reason first amplifiers with bipolar transistors will be presented. The power MOSFETs will return when switched-mode amplifiers are presented because of their excellent switching performance.

### 6.2.2 One transistor amplifiers

With both bipolar transistor types, several electronic functional building blocks can be realised. Many years ago all analogue electronics consisted of a combination of these configurations that were only built with discrete components. Presently most analogue electronics are realised with ICs called *Operational Amplifiers*. In order to better understand the working

principle of these integrated amplifiers, first the basic building blocks are explained in this section.

### 6.2.2.1 Emitter follower

The first circuit is the *emitter follower* from Figure 6.35. It is also called a *common collector* configuration as the collector is directly connected to the common power supply. The principle of the emitter follower is based on the current-amplification of the applied NPN transistor. It can for instance be used as the buffer amplifier in the cascaded passive first-order filter of Figure 6.18 to match a high impedance input with a low impedance output. In principle its operation is quite straightforward. The collector is connected to a positive voltage with in this example a value of +15 V. The emitter is connected to a more negative voltage via an emitter resistor. For the negative voltage a value of -15 V is chosen to be able to amplify alternating signals. The input voltage is applied on the base. When this input voltage has a value in between the positive and negative supply voltage, the base-emitter junction is in the conductive mode. Like with a Silicon diode, the conducting base-emitter junction has a threshold voltage  $V_{b-e,th}$  of  $\approx 0.6$  V. For that reason, the emitter voltage will be approximately 0.6 V more negative than the input voltage.

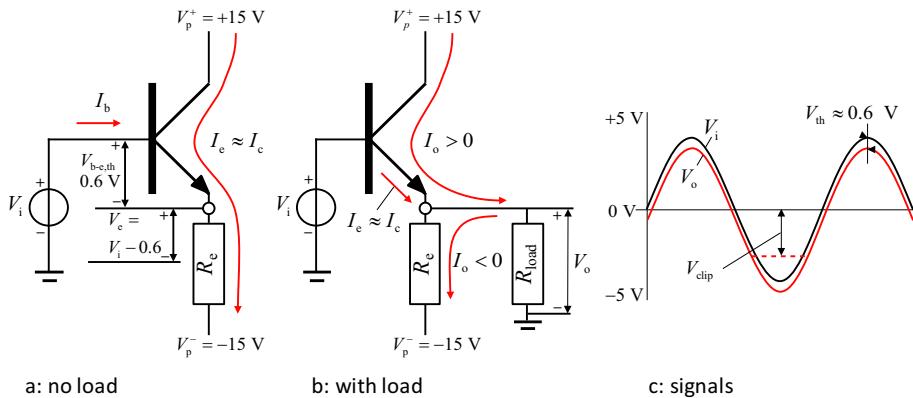
The difference between the emitter voltage and the negative supply voltage determines the emitter current  $I_e$  through the emitter resistor.

$$I_e = \frac{V_e - V_p^-}{R_e} = \frac{V_i - 0.6 + 15}{R_e} \quad (6.54)$$

This current is equal to the sum of the collector current and the base current:

$$I_e = I_b + I_c = I_c \left( \frac{1}{\beta} + 1 \right) \quad (6.55)$$

For high values of  $\beta$ , the collector current becomes almost equal to the emitter current and for this section the base current will further be neglected. The current in the emitter resistor is necessary to be able to deliver both positive and negative currents to a load, as shown in Figure 6.35.b, and is called the *idle current*. When the input voltage at the base is an alternating sinusoidal signal, like shown in Figure 6.35.c, then the emitter voltage follows the input voltage with a constant difference of  $\approx 0.6$  V. The external load will pull either a positive current or a negative current from the node between the emitter and the emitter resistor. The positive current is delivered by the transistor, on top of the current flowing into the emitter resistor.

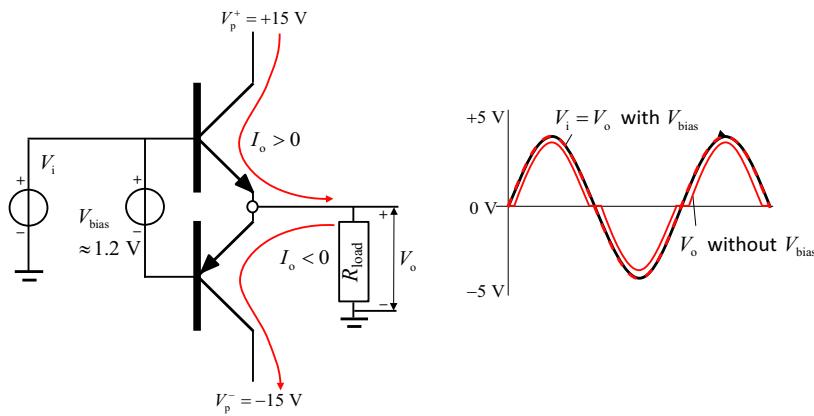


**Figure 6.35:** The NPN transistor emitter follower is a current amplifier used to achieve a low output impedance. The emitter voltage follows the input voltage that is applied to the base with a difference of  $\approx 0.6$  V. When a load is connected, the positive output current is delivered by increasing the emitter current of the transistor. The negative output current comes from the emitter resistor  $R_e$  by a reduction of the emitter current of the transistor. Ultimately the emitter current is zero and the maximum negative voltage is reached, showing a phenomenon called “clipping” at a level that is determined by the load in relation to the emitter resistor and the negative power supply voltage.

The negative current is delivered by the resistor by reducing the current from the transistor. This can be done until the emitter current is zero. The transistor is not capable of delivering a negative current and as a result the amplifier will show a *clipping* effect at a maximum negative level that can be calculated from the voltage divider that is determined by  $R_e$  and  $R_{\text{load}}$  with the negative supply voltage:

$$V_{\text{clip}} = V_p^- \frac{R_{\text{load}}}{R_e + R_{\text{load}}} \quad (6.56)$$

This single transistor emitter follower is only suitable in situations, where no large output currents are required. This limitation is both due to the mentioned clipping at the negative cycle and also because of the heat dissipation in the emitter resistor and transistor by the idle current, even when no signal is amplified. This *single-ended Class A* configuration is especially inefficient with high power amplifiers. For that reason a more symmetrical solution is often applied, where the negative current is delivered by a PNP transistor. This configuration is shown in Figure 6.36. The input voltage



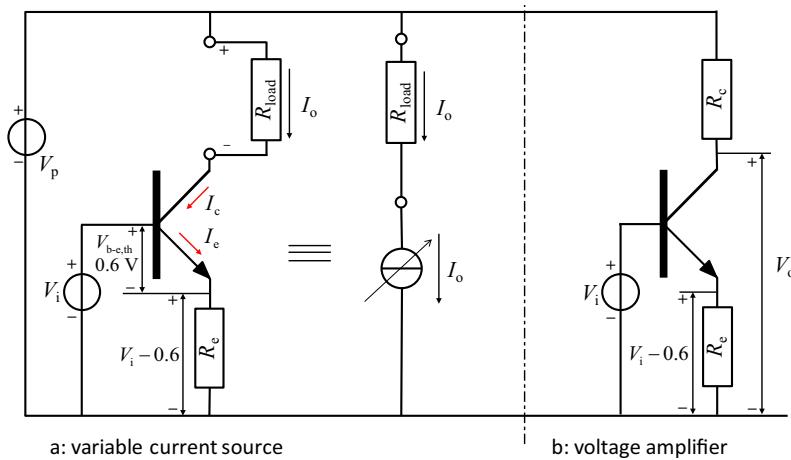
**Figure 6.36:** A symmetric push-pull emitter follower. The positive current is delivered by the NPN transistor and the negative current by the PNP transistor. A bias voltage is added to avoid non-linearity at the transition between a positive and a negative current, that would otherwise be caused by the threshold voltage of the base-emitter junction of the transistors.

is applied simultaneously to the base of both transistors. In the positive cycle of the signal the NPN transistor operates almost identical as in the single-ended configuration, but it only delivers current to the load, not into an emitter resistor. In the negative cycle the PNP transistor takes over the task to deliver the output current. This so called *Push-Pull class B* configuration has one little problem that is caused by the threshold voltage of the base-emitter junction of  $\approx 0.6\text{ V}$  per transistor. This gives a strong non-linearity at the transition between the positive and negative current, called *cross-over distortion*. Because the gain becomes almost zero at  $0\text{ V}$  input, feedback can not solve that non-linearity and it is necessary to reduce it by applying a bias voltage between the bases of both transistors. The tuning of this bias voltage is quite critical, due to the large temperature dependency of the base-emitter threshold voltage. In most cases a certain limited idle current is set running through both transistors and the bias voltage is temperature compensated to add robustness.

This configuration is called *Push Pull class AB*.

### 6.2.2.2 Voltage amplifier

Based on the findings with the emitter follower, two other configurations are shown in Figure 6.37, the variable current source and the voltage ampli-



**Figure 6.37:** Two configurations where the load is applied to the collector.

- a: With the variable current source, the collector current is only determined by the input voltage and the emitter resistor. It has a very high output impedance.
- b: A voltage amplifier is created from a variable current source, by inserting a resistor between the positive power supply and the collector. The voltage over the resistor is proportional to the current and the resistor value. With a high value for  $R_c$ , voltage amplification is obtained at the collector output.

fier. Both circuits belong to the same type and are called *collector follower* or *common emitter* configurations, because the output is taken from the collector and the emitter is either directly or almost directly connected to the common ground.

Like with the example of the emitter follower, the NPN transistor is used as the representative example. For the PNP transistor only the signs of the voltages and currents need to be reversed to get the same results.

The circuit of Figure 6.37.a. acts as a variable current source with a negative current and works as follows: As long as the voltage of the collector is kept at any value more positive than the voltage at the base, the emitter current is only determined by the input voltage and the emitter resistance like with the emitter follower. With a high value of the current-amplification ratio  $\beta$ , the base current can be neglected and the collector current becomes equal to the emitter current. As a consequence of this reasoning, the collector current is not dependent on the collector voltage, which is the defined behaviour of a current source.

The current source will prove to be very useful in the design of a differential

amplifier, but first it is used to create a single transistor voltage amplifier by including a collector resistor between the collector and the positive supply rail, as shown in Figure 6.37.b.

The current in the collector determines the voltage over the resistor resulting in an output collector voltage  $V_o$  equal to:

$$V_o = V_p - I_c R_c = V_p - I_e R_c = V_p - \frac{(V_i - V_{th}) R_c}{R_e} \quad (6.57)$$

For the AC part of the input signal the constant values of  $V_{th}$  and  $V_p$  don't matter, which means that this configuration gives an AC voltage amplification of:

$$G_a = \frac{V_o}{V_i} = -\frac{R_c}{R_e} \quad (6.58)$$

With an ideal transistor this gain is independent of the values of the resistors. The minus term means that the sign of the signal is inverted. Although this looks equal to a  $180^\circ$  phase shift this effect is not related to a phase delay but only with a sign reversal.

Like all components that were presented, also the transistor is not ideal. Its current-amplification ratio is not infinite nor constant and the characteristics are non-linear because of the base-emitter junction, as was shown for a diode in Figure 6.30. Furthermore, like the diode, also the transistor has its dynamic limitations, that will be dealt with after the section about the operational amplifier.

Finally, the emitter voltage does not follow the base voltage perfectly, but it shows a certain output resistance  $R_{s,e}$  with a value of approximately:

$$R_{s,e} = \frac{0.025}{I_C} \quad [\Omega] \quad (6.59)$$

while this value is also temperature dependent.

Due to this finite emitter impedance the amplification ratio of Equation (6.58) can not be made infinite by choosing  $R_e$  equal to zero. In that case the internal emitter output resistance  $R_{s,e}$  determines the maximum amplification. With for instance a idle current of 1 mA, this output resistance is 25  $\Omega$  and with a supply voltage of 20 V and a 10 k $\Omega$  collector resistance, keeping the collector voltage at about 10 V, a gain of 400 is achieved. This number is reduced by the not infinite current-amplification ratio and, because of the non-linear base-emitter threshold voltage to current ratio, this amplifier is very non-linear. In fact a larger emitter resistor linearises the amplifier at



the expense of a reduced gain. In the past a lot of attention of electronic circuit designers was spent in optimising the settings of these single transistor amplifiers in order to avoid the use of multiple costly transistors.

### 6.2.2.3 Differential amplifier

As the last configuration, before integrating them all in an operational amplifier, the differential amplifier is presented in Figure 6.38. In essence this configuration consists of two single transistor voltage amplifiers that are connected via their emitters. It was explained that a voltage amplifier consists of a variable current source that creates a voltage over the collector resistor. This current is determined by the emitter resistor and the difference between the emitter voltage and the negative supply voltage. With the differential amplifier however, the collector current is not only determined by this difference but also by the difference between both base voltages. To explain the functionality, the collector resistor values of both transistors are assumed identical,  $R_{c,1} = R_{c,2} = R_c$ . Also both emitter resistors are assumed identical,  $R_{e,1} = R_{e,2} = R_e$ , while a third emitter resistor  $R_{c,e}$  is added that carries the current of both transistors.

The input signals  $V_{i,1}$  and  $V_{i,2}$  are modelled to consist of a common part and a differential part.

$$V_{i,1} = V_{i,c} + V_{i,d} \quad \text{and} \quad V_{i,2} = V_{i,c} - V_{i,d} \quad (6.60)$$

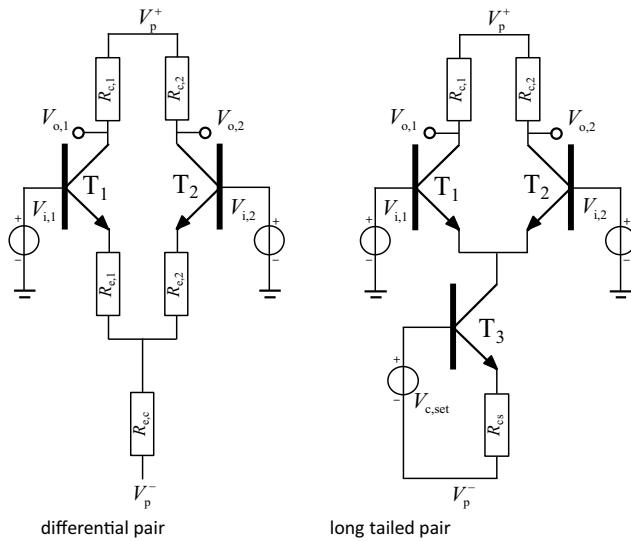
The common part of the signal is called the *common-mode signal* and this part results in an equal emitter current value  $I_{e,1} = I_{e,2}$  for both transistors at a value of half of the total current  $I_t$ . This total current is determined by total resistance to the negative supply voltage  $R_{e,t} = R_{e,c} + R_{e,1} \parallel R_{e,2} = R_{e,c} + 0.5R_e$ , because both emitter voltages are equal.

The part of the current in both transistors that is caused by the common-mode input signal can than be calculated as:

$$I_{e,1} = I_{e,2} = 0.5I_{e,t} = 0.5 \frac{V_{i,c} - 0.6}{R_{e,c} + 0.5R_e} \quad (6.61)$$

In this equation the threshold voltage  $V_{th} \approx 0.6$  V has been taken into account. The output voltage that is caused by the common-mode part of the input voltage is equal for both collectors, because of the equal collector resistances:

$$V_{o,1} = V_{o,2} = 0.5R_c \frac{V_{i,c} - 0.6}{R_{e,c} + 0.5R_e} \quad (6.62)$$



**Figure 6.38:** The differential amplifier consists of a pair of equal transistors with the emitters connected either directly or via a pair of emitter resistors.

- The differential amplification is determined by the collector resistors and the total resistance between the emitters while the common-mode amplification is determined by the collector resistors and the total resistance from the emitters to the negative supply voltage.
- With the high impedance current source  $T_3$ , the common-mode amplification is minimised.

The amplification of the differential part, the *differential-mode* signal, is only determined by the collector resistors and the series value of the resistors  $R_{e,1}$  and  $R_{e,2}$ . This is true, because for a differential signal the voltage at the connection point between both equal resistors does not change when one input voltage gets a positive  $dV_i$ , while the other input gets an equal negative  $dV_i$ . As a result, the common-mode current through  $R_{e,c}$  will be divided differently over both transistors in a ratio depending on the value of  $R_{e,1} = R_{e,2} = R_e$ . The current difference will then be:

$$dI_{e,1} = -dI_{e,2} = \frac{V_{i,1} - V_{i,2}}{2R_e} = \frac{dV_i}{R_e} \quad (6.63)$$

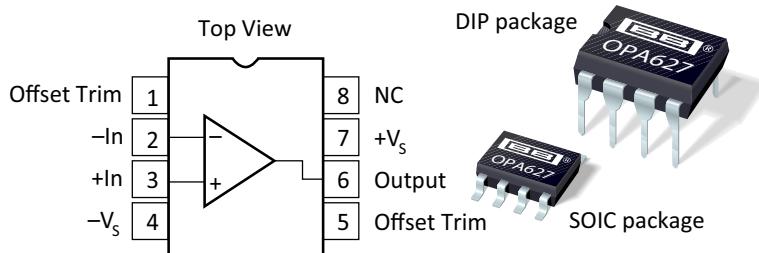
A positive difference between  $V_{i,1}$  and  $V_{i,2}$  gives a corresponding higher current in  $T_1$  and a decreased current in  $T_2$ . As a consequence, with the equal collector resistances, the part of the output voltage, caused by the

differential-mode part of the input voltage, becomes differential:

$$-V_{o,1} = V_{o,2} = R_c \frac{dV_i}{R_e} \quad (6.64)$$

With these relations it is possible to optimise the amplifier for a certain application. Generally the goal is to realise an ideal differential amplifier without any common-mode amplification. This is for instance useful in measurement systems, to avoid errors due to common-mode interference signals. This means, that the value of  $R_{e,c}$  needs to be maximised and the value for  $R_e$  needs to be minimised. This ultimately results in the configuration of Figure 6.38.b, where the common emitter resistor is replaced by a third transistor in the current source mode, with a fixed input voltage  $V_{c,set}$ , that determines the common current through  $R_{cs}$ . This configuration is called a *long tailed pair* because of the additional current source in the tail of the differential amplifier. In this example the two emitter resistors are omitted, in order to increase the gain to the maximum value possible. When linearity is needed they can be added also in this configuration like previously explained with the single transistor voltage amplifier.

This differential amplifier configuration gives two output voltages that can either be used directly by choosing one of the outputs, but mostly it is used integrated in an operational amplifier.



**Figure 6.39:** The symbol, pinning and housing of an operational amplifier.  
(Courtesy of Texas Instruments)

### 6.2.3 Operational amplifier

The operational amplifier is possibly the most important and successful example of an analogue IC. It is frequently used in measurement and sensor systems, power amplifiers and many other applications, both in consumer and professional equipment. Its success is largely based on its versatility, ease of use and reliable results. While it is an amplifier with a very high gain, it is always used with feedback and because of that situation, it involves the same dynamic issues on stability, that are known from any other feedback system. For that reason it is presented here quite thoroughly as in mechatronic systems all performance issues are related to dynamics and stability.

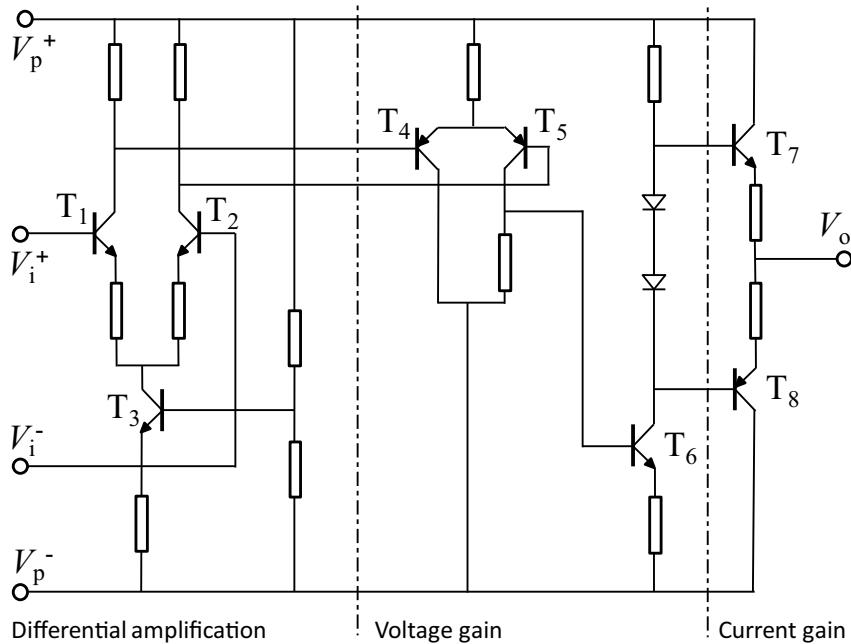
#### 6.2.3.1 Basic operational amplifier design

A basic operational amplifier consists of three functional stages as shown<sup>3</sup> in Figure 6.40.

The first stage, the differential input stage, is designed as a long tailed pair with current source  $T_3$ , that keeps the sum of the currents through  $T_1$  and  $T_2$  constant. This configuration maximises the ratio between the differential-mode gain and the common-mode gain, the *common-mode rejection ratio* (CMRR).

The second stage, the gain stage around  $T_4$ ,  $T_5$  and  $T_6$ , is a combination of a differential pair and a voltage amplifier. Without the resistors between the emitters, it creates a high differential gain with low linearity, which is

<sup>3</sup>To avoid confusion the voltage source for the power supply, input and output voltage is not drawn in the figure but all mentioned values are noted in respect to the reference which in general is ground or 0V from the power supply. This is quite common in electronic circuits.



**Figure 6.40:** Basic Operational amplifier with bipolar transistors, diodes and resistors. Three functional stages can be distinguished, a first stage that amplifies only the differential input voltage, a second stage with a maximum voltage gain and a third stage with a maximum current gain.

not a big issue<sup>4</sup> as an operational amplifier is uniquely used with feedback. With this gain the voltage difference from the first stage is multiplied. The last stage, the power output stage, has to be able to deliver the required current to the load. This is accomplished by a symmetrical power output stage in Class AB push-pull configuration. The necessary bias voltage is realised by the two diodes between the bases of T<sub>7</sub> and T<sub>8</sub> and solve the problem at the crossover between a positive and negative output current while they keep the transistors in just a conductive mode around a few mA of idle current.

<sup>4</sup>Neglecting the non-linearity is only allowed in non-critical situations. In reality, one should design any component in a control-loop as linear as possible in order to avoid residual errors.

### 6.2.3.2 Operational amplifier with feedback

An operational amplifier is always used with negative or positive feedback. Negative feedback is applied when a linear behaviour is needed and positive feedback is applied for special functions like the *Schmitt trigger* that will be used in Section 6.3 to design a pulse-width modulator for a switched-mode amplifier. When the term feedback is used in a linear system it generally means negative feedback, but in the section on active filters positive feedback is applied to create a suitable the transfer function.

For a non-inverting amplifier, applying negative feedback results in the representation as shown<sup>5</sup> in Figure 6.41. The circuit is drawn such that the classic feedback control loop is recognisable with a set point, an output and a differential stage, that compares the set point with the value fed back from the output. For an amplifier with an open-loop gain  $G_o$ , the closed-loop gain ( $G_a$ ) equals the complementary sensitivity function of the feedback circuit, as explained in Chapter 4 on motion control, and is given by:

$$G_a = \frac{V_o}{V_i} = \frac{G_o}{1 + G_o G_f} = \frac{1}{\frac{1}{G_o} + G_f} \quad (6.65)$$

Where the feedback gain  $G_f$  is given by the voltage divider:

$$G_f = \frac{R_1}{R_1 + R_2} \quad (6.66)$$

In an ideal operational amplifier  $G_o$  is infinite. This results in a closed-loop gain of:

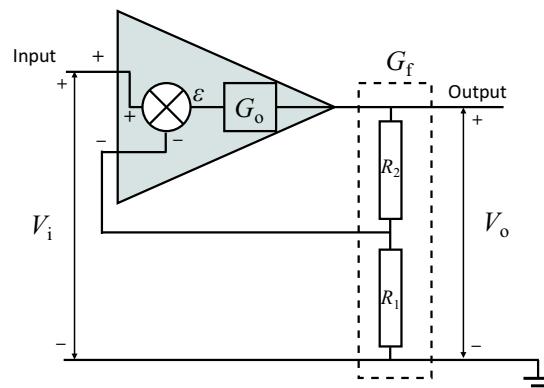
$$G_a = \frac{R_1 + R_2}{R_1} \quad (6.67)$$

This means that the amplification is only depending on the ratio of the resistors  $R_1$  and  $R_2$ . It also means that the difference in input voltages becomes zero at infinite  $G_o$ . In other words: “The output will do anything to make the voltage at the minus terminal, the feedback point, equal to the voltage on the plus terminal, the reference point”.

This understanding is used in most simplified designs of electronic circuits, like presented in the following section, where different amplifier and filter combinations are shown based on operational amplifiers.

---

<sup>5</sup>The power supply is omitted for simplicity in all further circuit diagrams with operational amplifiers. Unless otherwise mentioned a symmetrical power supply with an equal positive and negative supply voltage of sufficient magnitude is assumed to be present.



**Figure 6.41:** Operational amplifier is used with feedback.

### 6.2.4 Linear amplifiers with operational amplifiers

In this section several basic operational amplifier configurations will be presented using ideal operational amplifier characteristics:

- The open-loop gain is infinite.
- The input current is zero with an infinite input impedance.
- The output current is unlimited, with a zero output impedance.
- The common-mode amplification is zero.
- No dynamic-, offset-, power supply- or other limitations and deviations are present.

Even though in reality all these statements are not true, in many less critical applications general operational amplifiers behave sufficiently ideal to justify the use of the configurations in this section.

#### 6.2.4.1 Design rules

The following rules for designing linear circuits with operational amplifiers directly follow from the above mentioned understanding of the working principle of a feedback controlled operational amplifier and is used throughout all electronic circuits in this book, even when non-ideal amplifiers are considered:

1. When negative feedback is applied, the output does whatever it can to keep the negative input equal to the positive input ( $V^+ = V^-$ ).
2. The current in the circuit is not influenced by the input currents as these are zero.

These rules are applied in the following steps to determine the function of most operational amplifier circuits:

1. Check if negative feedback is present.
2. Calculate the voltage on the positive input as this is in most cases only determined by the input voltage and not by the feedback voltage.
3. Calculate the currents in the resistors and remaining voltages by applying rule 1 and 2 .

#### 6.2.4.2 Non-inverting amplifier

The non-inverting amplifier configuration is equal to the previously presented operational amplifier with feedback of Figure 6.41. It is drawn in its simplified form in Figure 6.42 with a special version in the second picture, which is the unity-gain non-inverting amplifier or buffer amplifier, that was used in the previous part about passive filters, to reduce the mutual interaction of the different filter sections.

Although the closed-loop gain is already determined previously, the validity of the rules and steps is verified by applying them also in this configuration. First it is checked that indeed negative feedback is applied, which is true via the voltage divider by  $R_1$  and  $R_2$ .

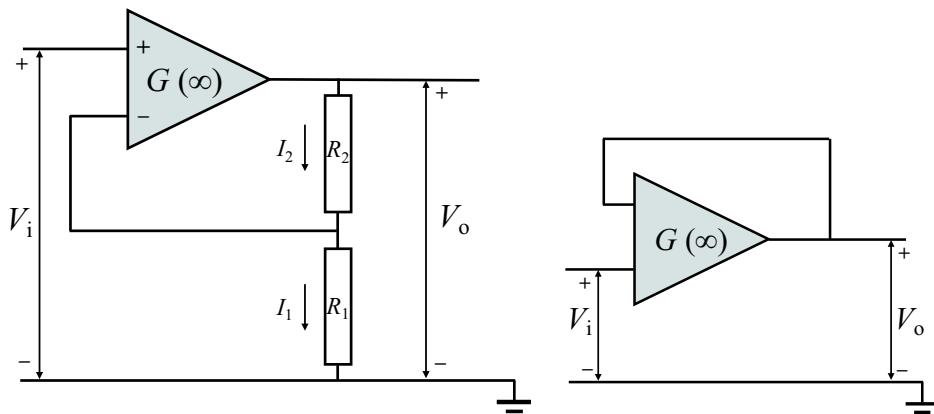
Then rule one is applied and the amplifier will do whatever it can to make voltage  $V^-$  at the negative input equal to the voltage at the positive input,  $V^+ = V_i$ .

With these conditions the current  $I_1$  in  $R_1$  should be:

$$I_1 = \frac{V_i}{R_1} \quad (6.68)$$

The current  $I_2$  in  $R_2$  is equal to  $I_1$  because of rule two, the input current of the amplifier is zero. This means that the output voltage as function of the input voltage becomes equal to:

$$\frac{V_o}{V_i} = G_a = \frac{I_1}{V_i} (R_1 + R_2) = \frac{R_1 + R_2}{R_1} \quad (6.69)$$

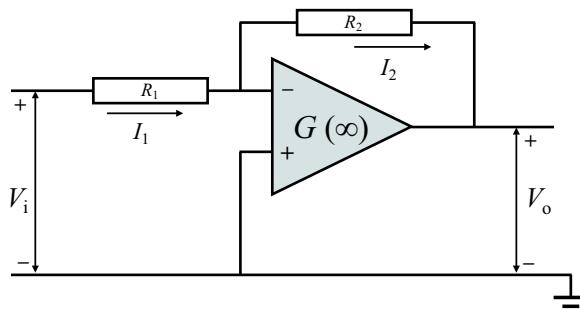


**Figure 6.42:** The non-inverting amplifier has a gain larger than one, determined by the voltage divider in the feedback path. When the output is directly connected to the minus input, giving a feedback gain of one, the unity-gain buffer amplifier is obtained.

This equation is equal to the previously obtained Equation (6.67) and the rules are verified.

When the resistors in the feedback path are chosen such, that  $R_1 = \infty$  and  $R_2 = 0$ , the gain  $G_a$  becomes equal to one. As a result, this configuration acts as a unity-gain buffer amplifier. A very important property of a non-inverting amplifier is the high input resistance of the circuit, which is only determined by the common-mode input resistance of the operational amplifier at the + input multiplied by the open-loop gain of the amplifier as the negative input is made to follow the positive input. This open-loop value of this resistance is already high and with feedback it is almost infinite even with a real operational amplifier. The differential input impedance does not play a role as the voltage on the - input is equal to the voltage on the + input. This configuration is especially useful when very weak sensor signals with a low voltage and high source impedance need to be amplified.

On the other hand, the output resistance of this amplifier is very low, because of the ideal operational amplifier characteristics. Even if there was a small output resistance, the feedback will reduce the internal output resistance by the total loop gain, according to the sensitivity function of the feedback system.



**Figure 6.43:** In an inverting amplifier the output has a different sign as the input.

### 6.2.4.3 Inverting amplifier

The inverting amplifier is shown in Figure 6.43. The amplification is calculated as follows.

According to the rules first the presence of negative feedback is checked, which is true via  $R_2$ . Then the positive input voltage is zero as it is connected to ground.

With rule one the output will take care that  $V^+ = V^- = 0$  and with this information the current  $I_1$  through  $R_1$  can be calculated by applying Ohms law:

$$I_1 = \frac{V_i - V^-}{R_1} = \frac{V_i}{R_1} \quad (6.70)$$

$I_2$  relates to the output voltage as follows:

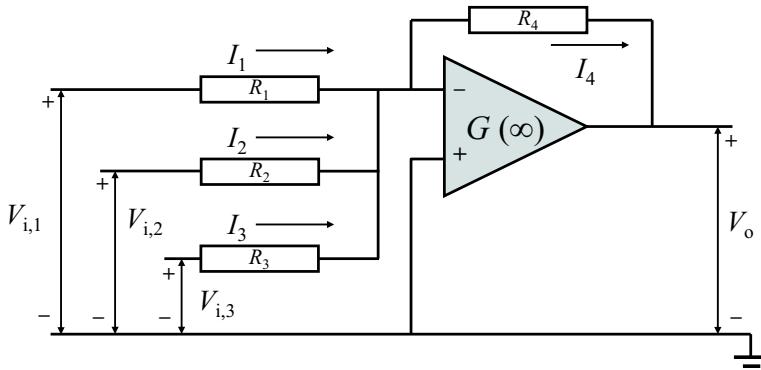
$$I_2 = \frac{V^- - V_o}{R_2} = -\frac{V_o}{R_2} \quad (6.71)$$

The current  $I_2$  through  $R_2$  is equal to  $I_1$ , because of rule two, resulting in the following closed-loop gain of the circuit:

$$I_1 = I_2 = \frac{V_i}{R_1} = -\frac{V_o}{R_2} \implies G_a = \frac{V_o}{V_i} = -\frac{R_2}{R_1} \quad (6.72)$$

The minus sign indicates, that the output signal is both amplified and inverted, as it has a different sign than the input signal. Contrary to the non-inverting amplifier, the closed-loop feedback gain can also be lower than one by choosing  $R_2 < R_1$ .

The main drawback of the inverting amplifier is its relatively low input resistance, that equals  $R_1$ . For that reason this configuration is applied in those situations, where the low input resistance gives no problems and the inverted functionality is needed. It is also useful, when signals have to be added or subtracted.



**Figure 6.44:** Adding signals is achieved by adding the currents of different inputs at the negative input, that acts like a virtual ground.

When  $R_1 = R_2 = R_3 = R$  the output signal is equal to the inverted sum of all input signals, amplified with a factor  $R_4/R$ .

#### 6.2.4.4 Adding and subtracting signals

A very powerful application of operational amplifiers is based on their capability to add and subtract signals in a configuration, as shown in Figure 6.44.

The calculation of the amplification is as follows. Again the rules are applied. There is negative feedback and the positive input voltage is zero and equal to the negative input voltage,  $V^+ = V^- = 0$ .

Because of rule two, no current can flow into the amplifier inputs and the currents through the resistors relate according to Kirchhoff's first law:

$$I_1 + I_2 + I_3 - I_4 = 0 \implies I_1 + I_2 + I_3 = I_4 \quad (6.73)$$

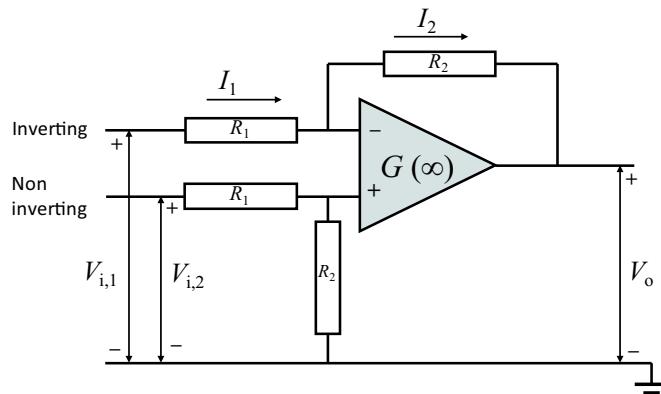
With Ohm's law these currents are calculated from the different voltage and resistor values:

$$\frac{V_{i,1}}{R_1} + \frac{V_{i,2}}{R_2} + \frac{V_{i,3}}{R_3} = -\frac{V_o}{R_4} \quad (6.74)$$

This results in the following output voltage:

$$V_o = G_{a,1}V_{i,1} + G_{a,2}V_{i,2} + G_{a,3}V_{i,3} = -\frac{R_4}{R_1}V_{i,1} - \frac{R_4}{R_2}V_{i,2} - \frac{R_4}{R_3}V_{i,3} \quad (6.75)$$

In the inverting and the adding configuration, the negative input acts like a *virtual ground*, because the output keeps the negative input at the same



**Figure 6.45:** Subtracting signals is achieved by combining a non-inverting and an inverting configuration and using both the positive and negative input of an operational amplifier.

level as the grounded positive input. This is the reason, that the signal at one input terminal is not influenced by the signal at another input terminal. This is especially useful, when signals are combined from sources that are so sensitive that other voltages on its output can cause interference by non-linearity.

This is different from the non-inverting amplifier.

When signals need to be subtracted, the circuit as shown in Figure 6.45 can be used. It is a combination of a non-inverting and an inverting amplifier configuration within one operational amplifier. The standard rules are applied also for this calculation of the amplification.

As the first step it is checked that negative feedback is applied via  $R_2$ . The next step is to calculate the voltage at the positive input and directly apply rule one:

$$V^- = V^+ = V_{i,2} \left( \frac{R_2}{R_1 + R_2} \right) \quad (6.76)$$

Using rule two, the relation between the output voltage  $V_o$ , the first input voltage  $V_{i,1}$  and  $V^-$  can be determined by using Ohm's law and Kirchhoff's first law:

$$I_1 - I_2 = \frac{V_{i,1} - V^-}{R_1} - \frac{V^- - V_o}{R_2} = 0 \quad (6.77)$$

Note the minus sign in  $I_2$  that is caused by the defined direction of the arrow for  $I_2$  in Figure 6.45.

With a little bit of algebra, this equation gives the relation for  $V^-$ :

$$V^- = V_o \frac{R_1}{R_1 + R_2} + V_{i,1} \frac{R_2}{R_1 + R_2} \quad (6.78)$$

When replacing  $V_-$  by the result of Equation (6.76), the following relation between the three voltages is obtained:

$$V_{i,2} \left( \frac{R_2}{R_1 + R_2} \right) = V_o \frac{R_1}{R_1 + R_2} + V_{i,1} \frac{R_2}{R_1 + R_2} \implies R_1 V_o = R_2 (V_{i,2} - V_{i,1}) \quad (6.79)$$

Resulting in:

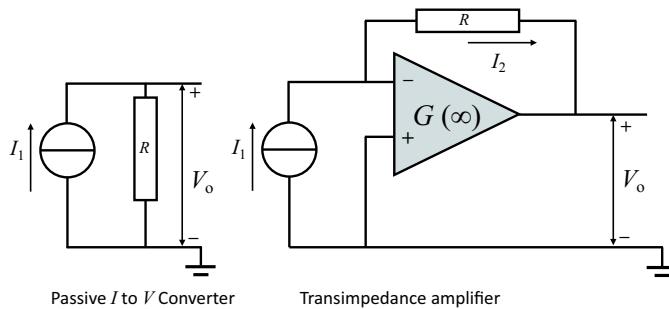
$$V_o = \frac{R_2}{R_1} (V_{i,2} - V_{i,1}) \quad (6.80)$$

It is important to be aware that the voltage divider with  $R_1$  and  $R_2$  after  $V_{i,2}$  with the same values as the resistors in the feedback circuit is necessary, because of the difference of amplification of a non-inverting and an inverting amplifier configuration. Sometimes the mistake is made to leave the resistors at the positive input away for simplification. This results in a higher gain for the non-inverted input signal, relative to the inverted input signal with a corresponding high common-mode amplification. In the ideal situation this common-mode amplification is zero when all resistances have an equal value.

A disadvantage of this single operational amplifier subtracting configuration is the difference in loading of the two input signals. The positive input has an input impedance equal to  $R_1 + R_2$ , while the negative input has an input impedance of  $R_1$  only. Further the current  $I_1$  is determined by both  $V_{i(1)}$  and  $V_{i(2)}$ , which means that  $V_{i(2)}$  influences the current from  $V_{i(1)}$ . This can cause problems when the source of  $V_{i(1)}$  has a voltage dependent impedance. In those cases, this has to be solved, for instance by inserting a separate unity-gain buffer amplifier between this source and the inverting input. In Chapter 8 the instrumentation amplifier will be introduced, a differential amplifier that uses two special buffer amplifiers, one for each input, that solves this issue in the most elegant way.

#### 6.2.4.5 Transimpedance amplifier

Several sensors that are applied in measurement systems give a current signal as function of the measurement parameter. The output impedance of these sensors is very high with a current source characteristic and the measurement current often needs to be converted into a voltage. Such a



**Figure 6.46:** A current to voltage converter creates an output voltage from the input current. A load resistor to ground acts like a passive  $I$  to  $V$  converter. With an operational amplifier one can create a converter that both adds energy to the signal and avoids a voltage over the current source.

converter is called a **current-to-voltage ( $I$  to  $V$ ) converter**, also named a **current controlled voltage source** or **transimpedance amplifier**.

In its simplest form a current to voltage converter is just a load resistor to ground as shown in the left drawing of Figure 6.46. The voltage over the resistor equals  $V_o = I_1 R$  according to Ohm's law.

This simple configuration has two drawbacks. First of all there is no energy amplification of the often very weak current signal and secondly the resulting voltage is also present over the input current source. When this current source is not ideal with a non-infinite output impedance, the voltage will influence the current.

To solve these problems, an operational amplifier can be used for the conversion as shown at the right drawing in Figure 6.46. The configuration is similar to the inverting amplifier but the input resistance is left away. This can be done because the impedance of the current source is in most cases much larger than the input resistor from a normal inverting amplifier and as they are connected in series the smaller value can be omitted.

Following the rules of an operational amplifier, the output of the amplifier will get a value such that the minus input voltage will become equal to the plus input voltage, being equal to 0 V ground level. With a zero current into the input of the operational amplifier the output voltage needs to be such that it creates a current  $I_2$  in the feedback resistor equal to the input current  $I_1$ .

The main differences with the simple resistor are clear. The virtual grounding of the input current at the minus input keeps the voltage over the input current source at zero Volt so even with a less perfect current source there is no problem. Furthermore the amplifier adds energy to the signal. Only

the sign is inverted and a positive current will result in a negative output voltage  $V_o = -I_1 R$ .

This active configuration acts like an impedance that the current has to pass through (trans) from the input to the output, for which reason this amplifier is called a transimpedance amplifier.

#### 6.2.4.6 Transconductance amplifier

In Chapter 5 it was shown that electromagnetic actuators require a controlled current in order to control the force. In such a case a converter is needed that converts the control voltage into a current with a current source characteristic.

Such a converter is called a voltage-to-current ( $V$  to  $I$ ) converter, also named a voltage-controlled current source or *transconductance* amplifier. The name transconductance is based on its inverted property when compared with the transimpedance amplifier where conductance is the inverse of impedance. Although in theory it is possible to create such a converter with a high ohmic series resistor between the voltage source and the load such a resistor would reduce the energy of the signal and for that reason a series resistor is not a realistic option when controlling actuators.

A transconductance amplifier can be created by introducing current measurement in the feedback loop of an operational amplifier as shown in Figure 6.47.

The left drawing shows a configuration where a small measurement resistor  $R_3$  is connected in series with the load that is connected to the output terminal of the amplifier. The current  $I_o$  creates a current sensing voltage  $V_{cs}$  over this resistor equal to  $V_{cs} = I_o R_3$ . This voltage is fed back to the negative input of the amplifier.

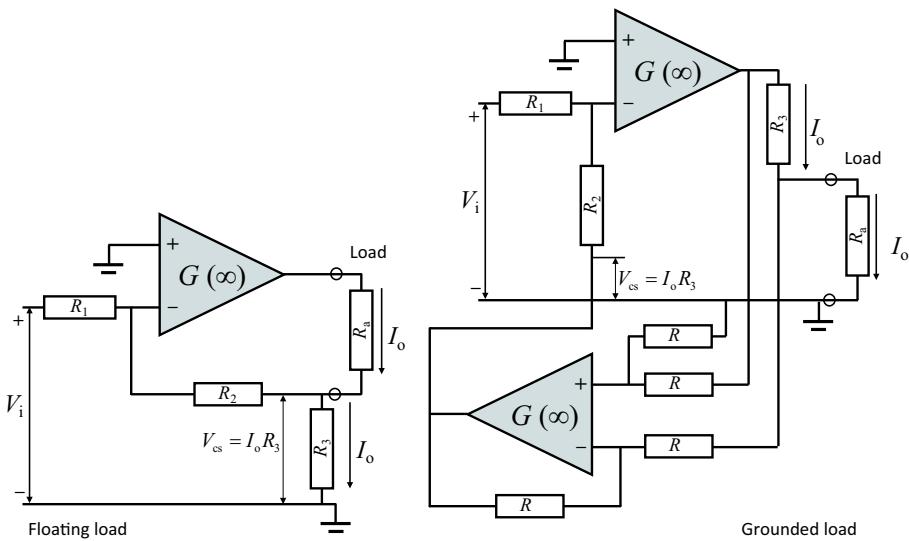
The amplifier obeys to rule one and does whatever it can to keep the negative input voltage equal to the positive input voltage, which is zero because the configuration is designed as an inverting amplifier. As a consequence of this design, the current sensing voltage  $V_{cs}$  over  $R_3$  will be kept at:

$$V_{cs} = -V_i \left( \frac{R_2}{R_1} \right) \quad (6.81)$$

To realise this voltage, the amplifier needs to supply an output current through the actuator and  $R_3$  at a value of:

$$I_o = \frac{V_{cs}}{R_3} = -V_i \left( \frac{R_2}{R_1 R_3} \right) \quad (6.82)$$

This means that the current through the load is only determined by the input voltage.



**Figure 6.47:** A voltage to current converter creates an output current that is determined by the input voltage. By measuring the output current with a small resistor and using that signal in the feedback loop a current source characteristic is realised. With a floating load this is simply achieved as shown in the left drawing. A grounded load requires a floating measurement resistor with a separate differential amplifier to create the current sensing voltage  $V_{cs}$  for current-feedback.

This simple configuration has one drawback in the fact that the load itself is not grounded at one of its terminals. This is called a *floating* load with a common-mode voltage relative to ground equal to the voltage over the current sensing resistor.

In case the load needs to be grounded the configuration at the right side of Figure 6.47 can be applied where the load and current sensing resistor are reversed. With this reversal the current sensing resistor is floating with a common-mode voltage that is equal to the voltage over the load. To cancel this common-mode voltage an additional subtracting amplifier is added that gives an output voltage to ground equal to the voltage over the current sensing resistor. The remaining part is equal to the floating-load configuration.

## 6.2.5 Active electronic filters

The presented amplifier configurations all used pure resistors to determine their closed-loop transfer function. In principle these resistors can be exchanged by complex impedances, in order to achieve a frequency dependent behaviour. Similar with passive filters, this could be done with capacitors and inductors, but the active nature of operational amplifiers creates the possibility, to avoid the large and expensive inductors and create filters of any kind and order with only capacitors and resistors. Also this field is extremely wide and requires a strict selection for this book. The examples chosen include simple first-order filters, an analogue PID controller and active second-order filters, as these are fully representative for the most frequently used configurations in mechatronic systems.

### 6.2.5.1 Integrator and first-order low-pass filter

For most of the filters, the inverting amplifier of Figure 6.43 is used as a starting point. The integrator and the related first-order low-pass filter are both shown in Figure 6.48, because they differ with only one resistor. Ohm's law is also valid for complex impedances, so the transfer function of an inverting amplifier according to Equation (6.72) can be written as follows:

$$G_a = -\frac{Z_2}{Z_1} \quad (6.83)$$

When adding a capacitor to one of these impedances, this transfer function becomes frequency dependent. If for instance  $R_2$  is infinite and a capacitor is placed parallel to it, the inverting amplifier becomes an inverting integrator with transfer function:

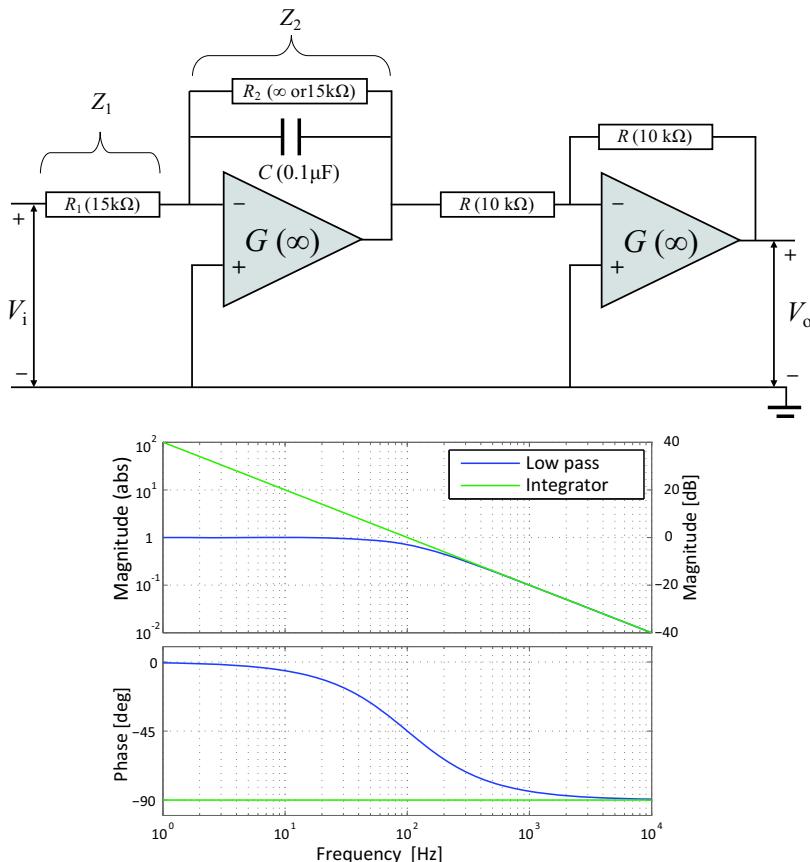
$$I^-(s) = -\frac{Z_2}{Z_1} = -\frac{1}{sR_1C}, \quad I^-(\omega) = -\frac{1}{j\omega R_1 C} \quad (6.84)$$

In order to compensate the minus sign, a second inverting amplifier has been added. This avoids the confusion of the phase of the integrator with the  $180^\circ$  phase of the inversion<sup>6</sup>. This results in the following integrator without inversion:

$$I(\omega) = \frac{1}{j\omega R_1 C} \quad (6.85)$$

---

<sup>6</sup>This additional inversion is often not necessary in complex filters with a multitude of inverting filter and amplification steps. In those situations eventually one additional inverter can be added if the resulting signal is still inverted.



**Figure 6.48:** A capacitor in the feedback path of an inverting amplifier creates an integrator, when the resistance of  $R_2$  is infinite. A finite value of  $R_2$  creates a first-order low-pass filter. The Bode plot shows both transfer functions using the indicated values of the passive elements. The second inverter with a gain of  $-1$  corrects the inversion by the first inverter.

In the Bode plot this non-inverting integrator gives a line with a slope of  $-1$ , that intersects the  $0 \text{ dB}$  level at the unity gain cross-over frequency  $\omega_0 = 2\pi f_0 = 1/R_1 C$  and a continuous phase of  $-90^\circ$  over the entire frequency band.

When the resistor  $R_2$  over  $C$  becomes smaller than infinite, it will limit the amplification at low frequencies and as a result a first-order low-pass filter is created. The transfer function is determined by calculating  $Z_2$  as the impedances of  $R_2$  and  $C$  in parallel and including the gain of  $-1$  by the

second inverter:

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2}{Z_1} = \frac{1}{R_1 \left( \frac{1}{R_2} + j\omega C \right)} = \frac{R_2}{R_1} \frac{1}{(1 + j\omega R_2 C)} \quad (6.86)$$

This represents a first-order filter with an RC-time constant  $\tau = R_2 C$  and a gain of  $R_2/R_1$  in the pass band below the corner-frequency  $\omega_0 = 2\pi f_0 = 1/\tau$ . At  $f_0$  the gain is  $-3$  dB and at higher frequencies the slope is  $-1$ , equal to the integrator. This is shown in the Bode plot for the indicated values of the passive elements, resulting in a pass band gain of one and  $f_0 = 100$  Hz.

Note that the unity-gain cross-over frequency  $\omega_0$  for the integrator and the corner-frequency  $\omega_0$  for the low-pass filter are equal in the Bode plot of Figure 6.48 because  $R_1$  and  $R_2$  are chosen to be equal for this example in order to fit into the same Bode Plot and enable comparing of the results.

Generally, even in normal inverting and non-inverting amplifiers, a small capacitor is almost always placed parallel to the feedback resistor over an operational amplifier. This is done for two reasons. First of all, it is never really useful in mechatronic systems, to amplify high frequencies above the range of interest, as at these frequencies many undesired dynamic eigenmodes are present in the mechanics. The second reason is related to the non-ideal properties of any electronic amplifier regarding its dynamic performance. More details will be presented after this section on filters, but it is already useful to recognise that a capacitor in the feedback path determines a differentiating action within the loop. This differentiation introduces a phase lead, that improves stability of the closed-loop system. This differentiating action of the capacitor is the result of the fact, that the current in a capacitor advances on the voltage and it is even better explained with the following example, the differentiator.

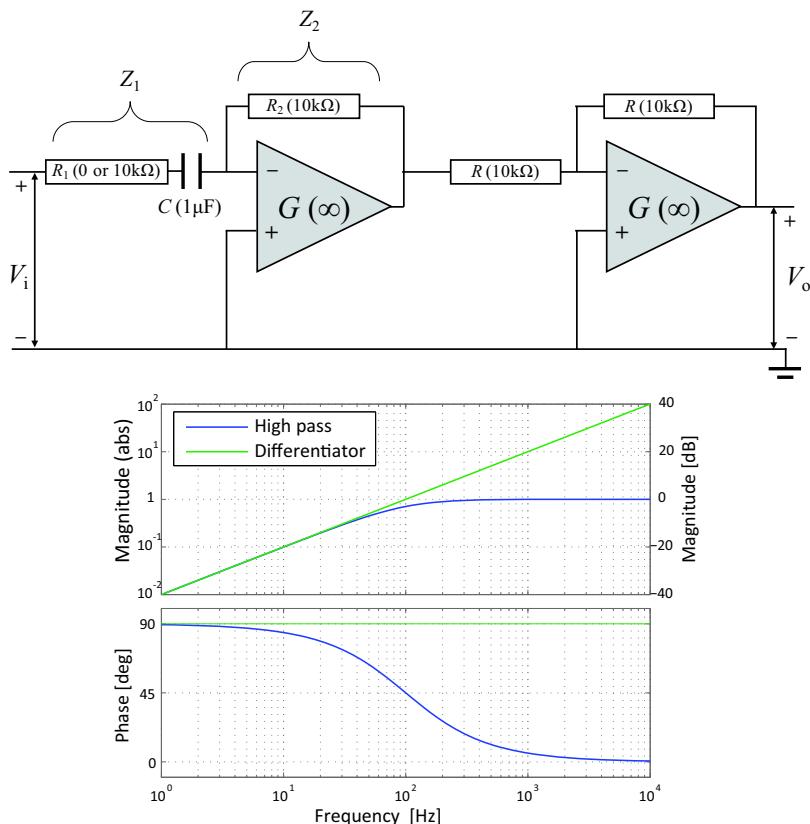
### 6.2.5.2 Differentiator and first-order high-pass filter

The differentiator and the almost identical first-order high-pass filter are created with the same inverting amplifier configuration, but in this case a capacitor is placed in the forward path as shown in Figure 6.49.

When the value of  $R_1$  is zero, a pure differentiator is created. When the gain of  $-1$  of the second inverter stage is included, the transfer function is:

$$D(s) = \frac{Z_2}{Z_1} = sR_2C, \quad D(\omega) = j\omega R_2C \quad (6.87)$$

In the Bode plot this results in a line with a slope of  $+1$ , that intersects the 0 dB level at the unity-gain cross-over frequency  $\omega_0 = 1/R_2C$  and a constant



**Figure 6.49:** A capacitor in the forward path of an inverting amplifier creates a differentiator, when the resistance of  $R_1$  is zero. A higher value of  $R_1$  creates a first-order high-pass filter. The Bode plot shows both transfer functions, using the indicated values of the passive elements.

phase lead of  $90^\circ$  over the entire frequency band.

Generally a pure electronic differentiator is impossible by definition, as the gain at infinitely high frequencies would need to become infinite. It is also far from practical in reality and measures should be taken to limit the gain of a differentiator at high frequencies. This is accomplished by adding a series resistor to  $R_1$  and this results in a first-order high-pass filter. The related transfer function, including the gain of  $-1$  of the second inverter, becomes:

$$F(\omega) = \frac{V_o}{V_i} = \frac{Z_2}{Z_1} = \frac{R_2}{\frac{1}{j\omega C} + R_1} = \frac{R_2}{R_1} \frac{1}{1 + \frac{1}{j\omega R_1 C}} = \frac{R_2}{R_1} \frac{j\omega R_1 C}{1 + j\omega R_1 C} \quad (6.88)$$

This represents a first-order filter with an RC-time constant  $\tau = R_1 C$  and a gain of  $R_2/R_1$  in the pass band above the corner-frequency  $\omega_0 = 1/\tau$ . At  $\omega_0$  the gain is -3 dB and at lower frequencies the slope is +1, equal to the differentiator. This is shown in the Bode plot for the indicated values of the passive elements, resulting in a pass band gain of one and  $\omega_0 = 100 \text{ rad/s}$ .

## 6.2.6 Analogue PID controller

Before the creation of fast digital controllers, operational amplifiers were used to realise the PID-control function. Figure 6.50 shows an example of such a circuit, that consists of a combination of the differentiator and integrator, that were presented before.

Before deriving the transfer function, first the three elements of control are examined using the values of the passive elements in the circuit. They will appear to be chosen according to the rules of thumb for a PID-controller, as described in Chapter 4 with a first order taming of the D-control action. Because this controller is meant to be used in the domain of mechanical engineering the frequency will be presented in Hertz and the magnitude in the Bode plot is absolute instead of in dB.

Starting at zero Hertz the impedance of the capacitors is infinite and the gain of the circuit will be infinite too, just as expected from an I-control system. At an increasing frequency the impedance of  $C_2$  will dominate the gain of the controller, because the impedance of  $C_1$  can still be neglected compared to the impedance of the parallel resistor  $R_3$ . As a result, the gain will show a  $-1$  slope in the Bode plot, conform the I-control action. At the integrator corner-frequency  $f_i = \omega_i/2\pi = 1/(2\pi R_2 C_2) \approx 20 \text{ Hz}$ , the impedance of  $C_2$  becomes smaller than the impedance of  $R_2$  and the negative slope will flatten out. At the differentiator corner-frequency  $f_d = \omega_d/2\pi = 1/(2\pi R_3 C_1) \approx 60 \text{ Hz}$ , the impedance of  $C_1$  becomes smaller than the impedance of  $R_3$  and the corresponding differentiating action will show a  $+1$  slope in the Bode plot, that is terminated (tamed PID-control) as soon as the impedance of  $C_1$  becomes smaller than  $R_1$ . This happens at the taming corner-frequency  $f_t = \omega_t/2\pi = 1/(2\pi R_1 C_3) \approx 600 \text{ Hz}$ .

As was presented in Chapter 4 on the PID controller, a differentiating action gives an additional gain at the cross-over frequency. In this example the second inverter is used to correct the proportional gain for this additional gain. The maximum phase lead of the controller occurs at the cross-over frequency of 200 Hz, where the gain is one. This value is chosen for this example. In reality in most cases a higher proportional gain is needed, to

compensate the low gain of the plant. The circuit can be adapted to this gain by changing the resistors of the second inverting amplifier.

### 6.2.6.1 Transfer function

The transfer function of the electronic PID-controller is derived in the frequency domain for reason of simplicity and starts with the same general transfer function for the inverting amplifier as with the previous filters, including the gain  $-g_i$  of second inverting amplifier:

$$C_{\text{pid}}(s) = \frac{V_o}{V_i} = g_i \frac{Z_2}{Z_1} = g_i \frac{\frac{1}{sC_2} + R_2}{\frac{1}{sC_1} + R_1 + \frac{1}{R_3}} \quad (6.89)$$

With shifting of the different terms the following equation is obtained:

$$C_{\text{pid}}(s) = g_i \left( \frac{1}{sC_2} + R_2 \right) \left( \frac{sC_1}{1 + sR_1C_1} + \frac{1}{R_3} \right) \quad (6.90)$$

At this point the relevant time constants of the controller are chosen according to the corner-frequencies, defined in the previous part. The first one,  $\tau_i$  defines the integrator corner-frequency  $f_i$ . The second one,  $\tau_d$  defines the differentiator corner-frequency  $f_d$  and the third one  $\tau_t$  defines the taming corner-frequency  $f_t$ .

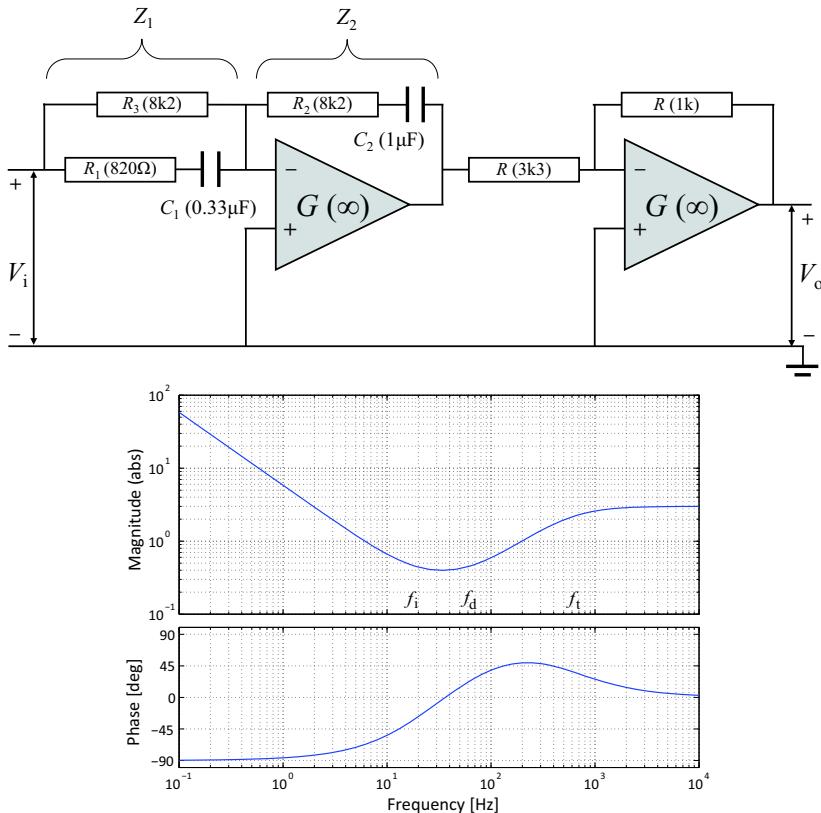
$$\begin{aligned} \tau_i &= \frac{1}{\omega_i} = \frac{1}{2\pi f_i} = R_2 C_2 \\ \tau_d &= \frac{1}{\omega_d} = \frac{1}{2\pi f_d} = (R_1 + R_3) C_1 \approx R_3 C_1 \\ \tau_t &= \frac{1}{\omega_t} = \frac{1}{2\pi f_t} = R_1 C_1 \end{aligned} \quad (6.91)$$

The approximation in  $\tau_d$  is in most cases allowed as  $R_1 \ll R_3$ .

By using these terms the following result can be obtained:

$$C_{\text{pid}}(s) = g_i \frac{R_2}{R_3} \left( \frac{1}{\tau_i s} + 1 \right) \left( \frac{sC_1 R_2}{1 + sR_1 C_1} + 1 \right) = g_i \frac{R_2}{R_3} \left( \frac{\tau_i s + 1}{\tau_i s} \right) \left( \frac{\tau_d s + 1}{\tau_t s + 1} \right) \quad (6.92)$$

The first term is a proportional term, the second term the integrator and the third term the tamed differentiator. As will be shown further on, the proportional term needs a correction factor, to become equal to the P-control gain  $k_p$



**Figure 6.50:** Analogue PID controller and its Bode plot. The indicated values of the resistors and capacitors are chosen for a PID controller, according to the rules of thumb for a mass control system, with a targeted unity-gain of the controller at the 200 Hz bandwidth. In reality the proportional gain should be adapted to the real gain of the plant at 200Hz.

In the Bode plot the frequency response of this controller is shown with the values as derived from the given components,  $\tau_i = 8.2 \cdot 10^{-3}$ ,  $\tau_d = 2.7 \cdot 10^{-3}$  and  $\tau_t = 2.7 \cdot 10^{-4}$ .

As presented in Chapter 4, the positive phase in the frequency band where the lead network is working, creates an effective damping in a motion control feedback configuration, like the servo system in a CD player. The integration at low frequencies reduces steady state errors, due to constant disturbing forces like gravity.

By tuning the values for the capacitors and resistors, the controller can be adapted to the plant by means of loop shaping.

### 6.2.6.2 Control gains

The control gains  $k_p$  for P-control,  $k_d$  for D-control and  $k_i$  for I-control, can be related to the time constants and passive component values of this controller. For this derivation the taming factor  $1/(\tau_t s + 1)$  is not taken into account as this represents just an additional pole, that is applied without impact on the standard control gains.

To determine the control gains, Equation (6.92) is written in the additive way by multiplication of the terms. This gives the following result, that corresponds with Equation (4.32) from Chapter 4:

$$C_{pid}(s) = g_i \frac{R_2}{R_3} \left( 1 + \frac{\tau_d}{\tau_i} \right) + g_i \frac{R_2}{R_3} \frac{1}{\tau_i s} + g_i \frac{R_2}{R_3} \tau_d s = \left( k_p + \frac{k_i}{s} + k_d s \right) \quad (6.93)$$

With this relation and applying the approximation in  $\tau_d$ , the proportional gain  $k_p$  is equal to:

$$k_p = g_i \frac{R_2}{R_3} \left( 1 + \frac{\tau_d}{\tau_i} \right) = g_i \frac{R_2}{R_3} \left( 1 + \frac{(R_1 + R_3)C_1}{R_2 C_2} \right) \approx g_i \left( \frac{R_2}{R_3} + \frac{C_1}{C_2} \right) \quad (6.94)$$

The I-control gain is equal to:

$$k_i = g_i \frac{R_2}{R_3} \frac{1}{\tau_i} = g_i \frac{R_2}{R_3} \frac{1}{R_2 C_2} = g_i \frac{1}{R_3 C_2} \quad (6.95)$$

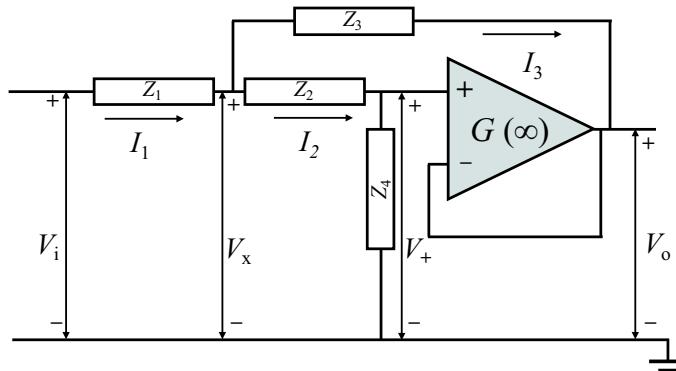
The D-control gain is equal to:

$$k_d = g_i \frac{R_2}{R_3} \tau_d = g_i \frac{R_2}{R_3} (R_1 + R_3) C_1 \approx g_i R_2 C_1 \quad (6.96)$$

Also in these expressions the approximation is based on the small value of  $R_1$  relative to  $R_3$ .

### 6.2.6.3 High speed PID control

Even though digital controllers have taken over many control actions in modern equipment, still one important advantage of analogue controllers over digital ones remains and that is the lack of delay due to sampling. In the digital world differentiating is done at any sampling moment by taking at least two samples, the present and the previous one, and dividing the difference over the sampling time. The result of this operation can only be supplied to the plant at the next sampling moment so at least one sample period later and as a result, part of the beneficial phase lead will be lost. One of the solutions is to use a model based controller, to predict the states, as explained in Chapter 4, but in unpredictable circumstances it is



**Figure 6.51:** Sallen-Key active second-order filter configuration with positive feed-back.

always better to avoid delays in the D-control part. This requirement has pushed sampling periods to ever smaller values, but in extremely fast control situations, like with a piezoelectric scanner for a video-rate Atomic Force Microscope, with bandwidths that exceed the Mega Hertz level, analogue controllers might still represent the best solution possible.

### 6.2.7 Higher-order electronic filters

In Section 6.1.3 on passive filters it was shown, that inductors are required, to achieve higher order electronic filters with an adequate transfer function. Especially in the low-frequency area, the required high values of the self-inductance lead to large sizes of the components and inductors in general are rather expensive to manufacture. In this section on electronic filters it will be shown with one example how, by using an operational amplifier, these filters can be created with only small size capacitors and resistors. In principle, these filters are designed to be applied with low power signals and not between amplifiers and actuators, so they can be used only in the path before the power amplifier.

The example that is presented here is a *Sallen-Key* filter named after R.P Sallen and E.L. Key, two engineers of MIT Lincoln laboratory in 1955. It is built around a non-inverting unity-gain amplifier, as shown in Figure 6.51, where the output signal is positively fed back to the input to create complex poles in the transfer function with imaginary parts.

The rules are applied again to calculate the transfer function. The positive input is equal to the voltage at the negative input and also equal to the

output voltage, as the operational amplifier acts as a unity-gain voltage follower:

$$V^+ = V_o \quad (6.97)$$

With this information current  $I_2$  is determined:

$$I_2 = \frac{V^+}{Z_4} = \frac{V_o}{Z_4} \quad (6.98)$$

The following step is to determine the voltage  $V_x$ :

$$V_x = V_o + I_2 Z_2 = V_o \left( 1 + \frac{Z_2}{Z_4} \right) \quad (6.99)$$

With this voltage  $I_1$  and  $I_3$  can be determined:

$$I_1 = \frac{V_i - V_x}{Z_1} = \frac{V_i - V_o \left( 1 + \frac{Z_2}{Z_4} \right)}{Z_1} I_3 = \frac{V_x - V_o}{Z_3} = \frac{V_o \left( 1 + \frac{Z_2}{Z_4} - 1 \right)}{Z_3} = V_o \frac{Z_2}{Z_3 Z_4} \quad (6.100)$$

With Kirchhoff's first law on currents, the sum of all currents at the common node of  $Z_1$ ,  $Z_2$  and  $Z_3$  equals zero. This means with the defined directions according to the arrows:

$$I_1 - I_2 - I_3 = 0 \quad (6.101)$$

Using the previously found equations for the different currents, with some algebra the following generic transfer function of this configuration is derived:

$$\frac{V_o}{V_i} = F = \frac{1}{\frac{Z_1 Z_2}{Z_3 Z_4} \frac{Z_1 + Z_2}{Z_4} + 1} \quad (6.102)$$

With this equation the transfer function of the second-order low-pass filter of Figure 6.52 will be determined as the first example.

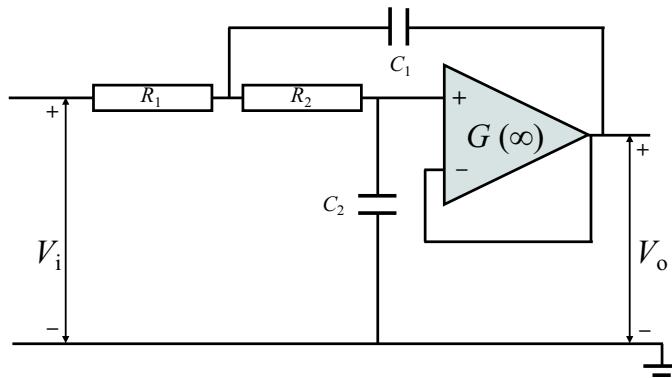
### 6.2.7.1 Second-order low-pass filter

The following values for the impedances are chosen for this example:

$$Z_1 = Z_2 = R_1 = R_2 = R, \quad Z_3 = \frac{1}{sC_1}, \quad Z_4 = \frac{1}{sC_2} \quad (6.103)$$

This gives as transfer function of the filter:

$$\frac{V_o}{V_i} = F(s) = \frac{1}{R^2 C_1 C_2 s^2 + 2RC_2 s + 1} \quad (6.104)$$



**Figure 6.52:** Sallen-Key active second-order low-pass filter.

This is the transfer function of a second-order low-pass filter, with the following values for the corner-frequency and damping:

$$\omega_0 = 2\pi f_0 = \frac{1}{R\sqrt{C_1 C_2}}, \quad \zeta = RC_2 \omega_0 = \sqrt{\frac{C_2}{C_1}}, \quad Q = \frac{1}{2\zeta} = \frac{1}{2}\sqrt{\frac{C_1}{C_2}} \quad (6.105)$$

It appears that the quality-factor  $Q$  can be made very high by relatively increasing the value of the positive feedback capacitor and a high damping can be made with a low value of the positive feedback capacitor. This clearly shows the effect on the imaginary part of the poles by the positive feedback.

### 6.2.7.2 Second-order high-pass filter

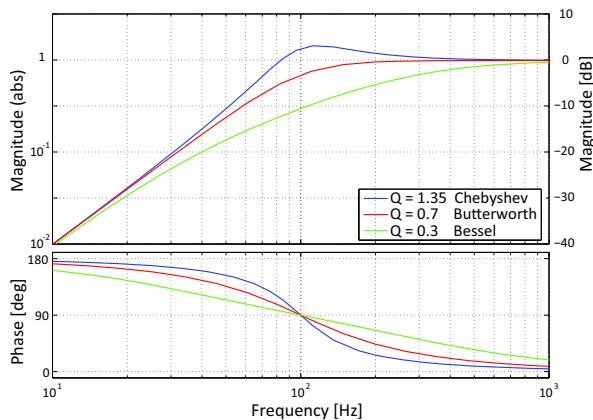
By exchanging the capacitors and resistors,  $R_1$  with  $C_1$  and  $R_2$  with  $C_2$ , a high-pass filter is obtained and it is left up to the reader to check the derivation of the following corresponding transfer function:

$$\frac{V_o}{V_i} = F(s) = \frac{R^2 C_1 C_2 s^2}{R^2 C_1 C_2 s^2 + 2RC_2 s + 1} \quad (6.106)$$

In this case the same values for  $\omega_0$ ,  $f_0$ ,  $\zeta$  and  $Q$  apply.

### 6.2.7.3 Different types of active filters

By choosing suitable values of the resistors and capacitors, electronic filters can be created with almost any frequency and level of damping, without the need for large inductors. They can be combined at will, to realise higher



**Figure 6.53:** Three typical characteristics of a second-order high-pass filter with the same corner-frequency but different levels of damping. For the Butterworth characteristics the  $-3\text{ dB}$  bandwidth is equal to the corner-frequency of 100Hz of this example. The Chebyshev characteristics result in a  $-3\text{ dB}$  bandwidth at 70 Hz with the steepest slope below the pass band. The Bessel characteristics result in a  $-3\text{ dB}$  bandwidth at 200 Hz and a very gradual slope below the pass band.

order filters of any slope in the attenuation band.

To illustrate this with an example, Figure 6.53 shows three typical, often used characteristics of a second-order high-pass filter with the same corner-frequency of 100 Hz, but with a different amount of damping.

When the damping has a  $Q$  of 0.7, the resulting lack of resonance and a magnitude of  $-3\text{dB}$  at the corner-frequency gives the filter *Butterworth* characteristics, named after the British physicist Stephen Butterworth (1885 –1958), who invented this filter type. This characteristic is often seen as the most ideal filter type, although the slope in the attenuation band, just below the  $-3\text{ dB}$  frequency determining the bandwidth, is not very steep. This steepness of the slope at the  $-3\text{ dB}$  frequency can be improved with a higher  $Q$  level. For instance a value of  $Q = 1.35$  results in the *Chebyshev* characteristics, named after the Russian mathematician Pafnuty Lvovich Chebyshev (1821 – 1894), because the characteristics are derived from his polynomials. The Chebyshev characteristic shows a ripple of  $+3\text{ dB}$  in the pass band, but also a more steep slope below the  $-3\text{ dB}$  frequency.

On the other hand, when phase characteristics need to be as gradual as possible, like in loudspeaker crossover filters, a lower value of  $Q$  is often preferred, like with the shown Bessel filter with  $Q = 0.3$ , named after the German astronomer and mathematician Friedrich Wilhelm Bessel (1784 –

1846). The Bessel characteristic shows a constant *group delay* in the pass band, representing a linear phase to frequency relationship  $d\phi/df$ .

The same characteristics can be shown for low-pass filters and these examples are only a limited set of the entire range of different filter configurations, that are designed for very specific purposes. Their design belongs to the domain of the specialist and with digital control even more transfer functions can be realised, including variable delay and other tricks. Nevertheless the mentioned configurations are frequently applied and already very suitable for the design of general mechatronic systems.

### 6.2.8 Ideal and real properties of operational amplifiers

All the previously shown configurations assumed ideal characteristics of the applied operational amplifier. One could wonder, why thousands of different types of these universal building blocks are designed and manufactured, when one would be sufficient. The reason is that like all components, also the operational amplifier is not ideal and the more it needs to approach the ideal behaviour, the higher the cost will be. In this section, different characteristics will be presented that limit the performance of the circuits with operational amplifiers. In most cases these limitations can be neglected, but it is important to be aware of them for the more critical applications with requirements, that are impaired by these limitations.

In Figure 6.54 an overview is given of the real characteristics of a typical high performance operational amplifier. The parameters, with their influence on the performance, will be explained separately. It will become clear that most limitations result from the basic building blocks of an operational amplifier, the transistors, diodes, resistors and capacitors.

First the dynamic parameters will be presented in a more thorough way, because of their relatively dominant influence on the functionality and stability of mechatronic systems. This is followed by a reduced presentation of the less impairing limitations.

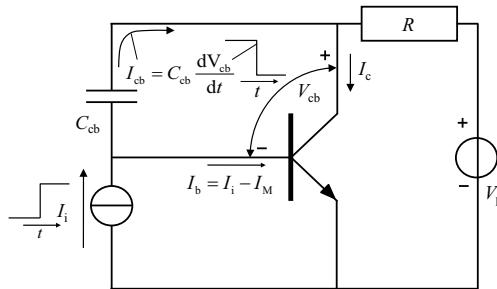
#### 6.2.8.1 Dynamic limitations

The most important limitation of all is determined by the finite gain and speed of an operational amplifier. Already at the presentation of a differentiator it was mentioned, that an infinite gain at infinite frequencies is impossible. This speed limitation is caused mainly by the parasitic capacitor over the PN-junctions of the applied transistors. Especially the parasitic

At  $T_A = +25^\circ\text{C}$ , and  $V_S = \pm 15\text{V}$ , unless otherwise noted.

PARAMETER	CONDITIONS	OPA627			UNITS
		MIN	TYP	MAX	
<b>OFFSET VOLTAGE<sup>(1)</sup></b>					
Input Offset Voltage AP, BP, AU Grades		40	10	250	$\mu\text{V}$
Average Drift AP, BP, AU Grades		100	0.4	0.8	$\mu\text{V}/^\circ\text{C}$
Power Supply Rejection	$V_S = \pm 4.5 \text{ to } \pm 18\text{V}$	0.8	2	2	$\mu\text{V}/^\circ\text{C}$
		120			dB
<b>INPUT BIAS CURRENT<sup>(2)</sup></b>					
Input Bias Current Over Specified Temperature SM Grade	$V_{CM} = 0\text{V}$	1	5	5	pA
Over Common-Mode Voltage	$V_{CM} = 0\text{V}$		1	1	nA
Input Offset Current Over Specified Temperature SM Grade	$V_{CM} = \pm 10\text{V}$	50	50	50	nA
	$V_{CM} = 0\text{V}$	1	5	5	pA
	$V_{CM} = 0\text{V}$	0.5	1	1	pA
			50	50	nA
					nA
<b>NOISE</b>					
Input Voltage Noise Noise Density: $f = 10\text{Hz}$		15	40	40	$\text{nV}/\sqrt{\text{Hz}}$
$f = 100\text{Hz}$		8	20	20	$\text{nV}/\sqrt{\text{Hz}}$
$f = 1\text{kHz}$		5.2	8	8	$\text{nV}/\sqrt{\text{Hz}}$
$f = 10\text{kHz}$		4.5	6	6	$\text{nV}/\sqrt{\text{Hz}}$
Voltage Noise, BW = 0.1Hz to 10Hz		0.6	1.6	1.6	$\mu\text{V}_\text{p-p}$
Input Bias Current Noise Noise Density, $f = 100\text{Hz}$		1.6	2.5	2.5	$\text{fA}/\sqrt{\text{Hz}}$
Current Noise, BW = 0.1Hz to 10Hz		30	60	60	$\text{fAp-p}$
<b>INPUT IMPEDANCE</b>					
Differential		$10^{13} \parallel 8$			$\Omega \parallel \text{pF}$
Common-Mode		$10^{13} \parallel 7$			$\Omega \parallel \text{pF}$
<b>INPUT VOLTAGE RANGE</b>					
Common-Mode Input Range Over Specified Temperature		$\pm 11$	$\pm 11.5$		V
Common-Mode Rejection	$V_{CM} = \pm 10.5\text{V}$	$\pm 10.5$	$\pm 11$		V
		106	116		dB
<b>OPEN-LOOP GAIN</b>					
Open-Loop Voltage Gain Over Specified Temperature SM Grade	$V_O = \pm 10\text{V}, R_L = 1\text{k}\Omega$	112	120	120	dB
	$V_O = \pm 10\text{V}, R_L = 1\text{k}\Omega$	106	117	117	dB
	$V_O = \pm 10\text{V}, R_L = 1\text{k}\Omega$	100	114	114	dB
<b>FREQUENCY RESPONSE</b>					
Slew Rate: OPA627 OPA637	$G = -1, 10\text{V Step}$	40	55		$\text{V}/\mu\text{s}$
Settling Time: OPA627 0.01%	$G = -4, 10\text{V Step}$	100	135		$\text{V}/\mu\text{s}$
0.1%	$G = -1, 10\text{V Step}$		550		ns
OPA637 0.01%	$G = -1, 10\text{V Step}$		450		ns
0.1%	$G = -4, 10\text{V Step}$		450		ns
Gain-Bandwidth Product: OPA627 OPA637	$G = -4, 10\text{V Step}$		300		ns
Total Harmonic Distortion + Noise	$G = 1$		16		MHz
	$G = 10$		80		MHz
	$G = +1, f = 1\text{kHz}$		0.00003		%
<b>POWER SUPPLY</b>					
Specified Operating Voltage Operating Voltage Range		$\pm 4.5$	$\pm 15$		V
Current			$\pm 7$	$\pm 18$	V
				$\pm 7.5$	mA
<b>OUTPUT</b>					
Voltage Output Over Specified Temperature	$R_L = 1\text{k}\Omega$	$\pm 11.5$	$\pm 12.3$		V
Current Output		$\pm 11$	$\pm 11.5$		V
Short-Circuit Current		$\pm 35$	$\pm 45$		mA
Output Impedance, Open-Loop	$V_O = \pm 10\text{V}$		$+70/-55$		mA
			55		$\Omega$
<b>TEMPERATURE RANGE</b>					
Specification: AP, BP, AM, BM, AU SM		-25	+85		$^\circ\text{C}$
Storage: AM, BM, SM AP, BP, AU		-55	+125		$^\circ\text{C}$
$\theta_{J-A}$ : AM, BM, SM AP, BP AU		-60	+150		$^\circ\text{C}$
		-40	+125		$^\circ\text{C}$
			200		$^\circ\text{C}/\text{W}$
			100		$^\circ\text{C}/\text{W}$
			160		$^\circ\text{C}/\text{W}$

**Figure 6.54:** Characteristics of a typical high performance operational amplifier.  
An extracted version of the original data sheet of the the OPA 627 of Burr Brown. (Courtesy of Texas Instruments)



**Figure 6.55:** Speed limitation due to the collector base capacitor  $C_{cb}$ . A voltage swing at the collector causes a current in  $C_{cb}$ , that has to be supplied by the input voltage. With a voltage amplifier, this is perceived at the input as a Miller capacitor with a larger value than  $C_{cb}$ .

capacitor  $C_{cb}$  between the collector and base is important, as it reduces the current-amplification at higher frequencies. The effect of this capacitor on a stepwise change in the input current is shown in Figure 6.55).

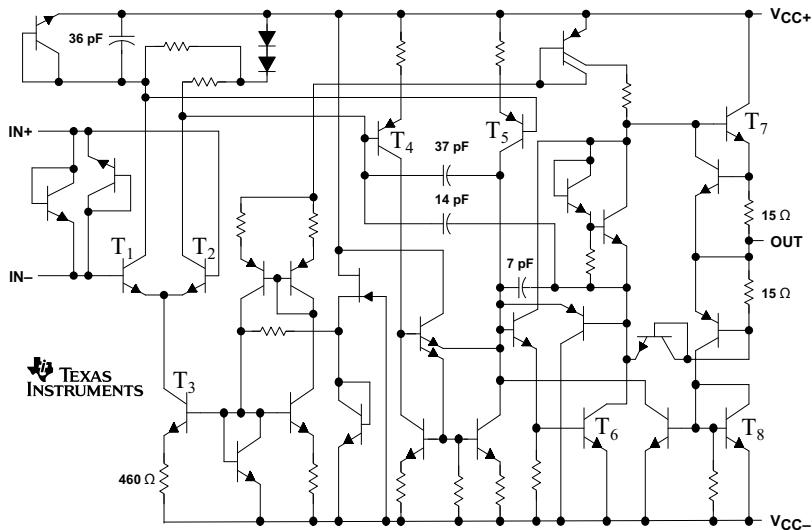
This input current ( $I_i$ ) would normally be equal to the base current ( $I_b$ ) and cause a proportional change in the collector current ( $I_c$ ). The collector current will cause a proportional drop of the collector voltage over resistor  $R$ , that induces a current ( $I_{cb}$ ) in  $C_{cb}$ , that in its turn reduces the current in the base:

$$I_b = I_i - I_{cb} = I_i - C_{cb} \frac{dV_{cb}}{dt} \quad (6.107)$$

To achieve a high speed, additional base current is necessary to compensate this effect depending on the capacitor value. Because of the relation of the current in  $C_{cb}$  with the voltage swing at the collector, this parasitic capacitor is perceived at the input as a much larger capacitor, the *Miller capacitor*  $C_M$ .

$$C_M = C_{cb}(G_v + 1) \quad (6.108)$$

where  $G_v$  is equal to the voltage amplification ratio of the transistor. With an emitter follower this capacitor can be neglected but with a voltage follower with an amplification of 100 even a value of 4 pF of a normal signal transistor becomes significant. The resulting Miller capacitor of 400 pF, combined with a practical input resistor of 10 kΩ, already limits the bandwidth of the single transistor voltage amplifier to  $f = 1/(2\pi RC) = 40$  kHz.

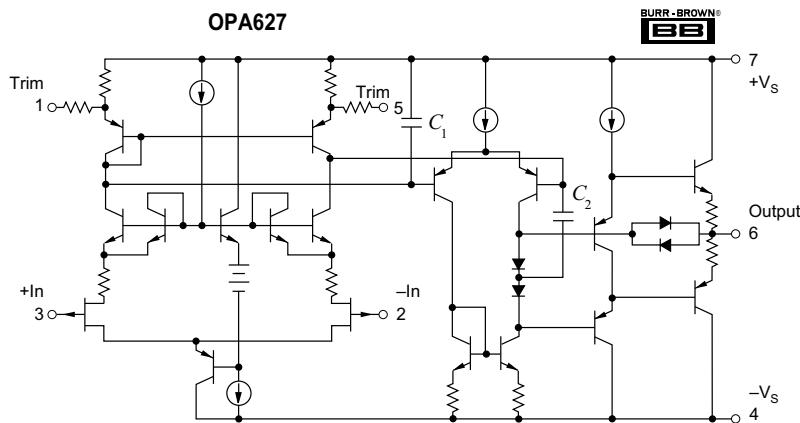


**Figure 6.56:** Schematic circuit drawing of the NE5532, a general purpose audio amplifier with bipolar transistor inputs. The transistors from Figure 6.40 are made recognisable and several more transistors are added for controlling the internal currents and voltages and for protection like the current protection transistors between T<sub>7</sub> and T<sub>8</sub>. The dominant pole is created by a feedback capacitor of 37 pF around the high gain stage by T<sub>4</sub> and T<sub>5</sub>. The other capacitors are added for further fine-tuning the open-loop frequency response.  
(Courtesy of Texas Instruments)

### Open-loop gain and frequency response

A real operational amplifier contains many transistors to control the internal currents and voltages and to protect the operational amplifier from overload. As examples the schematic circuit diagram of two real amplifiers, optimised for different purposes, are shown in Figure 6.56 and Figure 6.57. The NE5532 is a general purpose audio operational amplifier, designed for stability and low cost. The OPA627 is designed as a precision amplifier with Field Effect Transistors (FET) at the input to reduce the input currents. Like a MOSFET, a FET uses an electric field to control the current but the Gate is not insulated by an oxide but by another PN-junction.

The integration of many transistors in an operational amplifier results by definition in several parasitic capacitances. These capacitances cause problems in the dynamic performance, because they determine RC-times (poles) in the transfer function with the associated phase lag of 90° at  $\omega =$

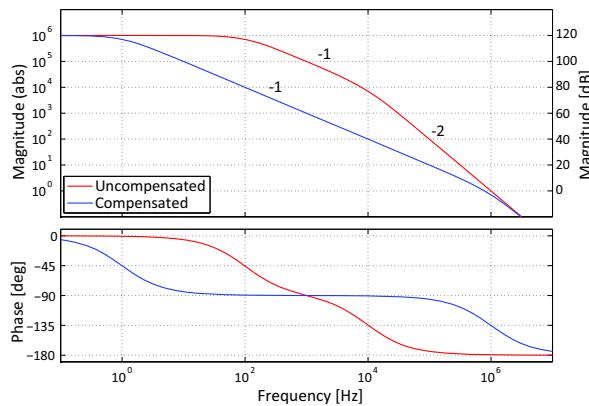


**Figure 6.57:** Schematic circuit drawing of the OPA627, a high precision amplifier with Field Effect Transistor (FET) inputs to reduce the input current. The dominant pole is created by increasing the pole of the parasitic capacitors around the large gain amplifier stage between  $C_1$  and  $C_2$ . (Courtesy of Texas Instruments)

$2\pi f = 1/RC$  for each capacitor with its source resistance. This effect is even increased with the collector-base capacitance of each transistor as explained with the Miller capacitor. When two or more of these first-order poles are located in the frequency range where the gain is larger than one, the closed-loop transfer function becomes marginally stable or even unstable when some additional phase lag is introduced by a third pole.

To illustrate this dynamic effect, an example open-loop frequency response with marginal stability is shown in Figure 6.58. The red line shows the frequency response of an amplifier with an open-loop gain of  $10^6$  and two poles below 1 MHz. The first pole is located at 100 Hz and the second pole at 10 kHz, where the gain is still  $10^4$ . The phase margin approaches zero even below 1MHz, where still some gain is present. This can create problems when the closed-loop gain is  $< 100$  as a phase margin of approximately  $45^\circ$  is preferred for a well damped behaviour.

This problem can be solved by limiting the open-loop gain and creating one dominant pole at a lower frequency. This causes the amplification to go down with a slope of  $-1$  from that frequency. The pole can be realised by placing an internal larger capacitor in the part of the operational amplifier, that is most prone to parasitic poles, like those related to the Miller capacitances in the high voltage gain stage. This capacitor can be placed parallel to an existing parasitic collector-base capacitance to move the main pole to



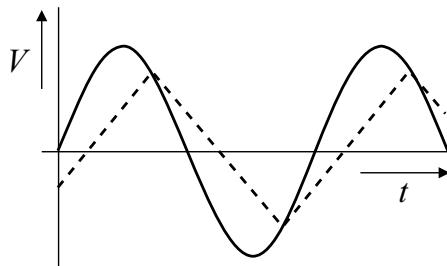
**Figure 6.58:** Poles in the transfer function of an operational amplifier can be cancelled by one dominant pole. This example shows two poles in the uncompensated situation, one at 100 Hz and one at 10 kHz. It is clear that the phase margin past 100 kHz is very small. With the dominant pole at 1 Hz in the compensated situation the phase margin becomes  $45^\circ$ .

a lower frequency, like  $C_2$  in the OPA627. But also a feedback capacitor can be used around the total high gain amplifier stage, like in the NE5532. The indicated feedback capacitor of 37 pF cancels one of the parasitic poles inside the loop and creates a new one, because it gives a differentiating action inside the loop (a zero).

By placing this pole at a sufficiently low frequency, the total gain of the amplifier crosses the unity-gain level before other poles have a significant impact on the phase. With this method, the circuit with an operational amplifier is stable at any closed-loop feedback configuration with a phase margin of  $45^\circ$  at the unity-gain cross-over frequency of 1 MHz for this example, as shown with the blue line in Figure 6.58. Because of the effect, that the first pole is moved to the left and the second pole to the right in the Bode plot, this method is called *pole-splitting*.

### Gain-bandwidth product

The creation of only one dominant pole in the frequency range where the gain is larger than one is reflected by a continuous  $-1$  slope in the frequency response above the corner-frequency that is determined by this pole. As a consequence the product of the gain and the frequency is constant at any frequency between the corner-frequency and the unity-gain cross-over fre-



**Figure 6.59:** The slew-rate of an operational amplifier limits the speed of change of the output voltage. At high frequencies the output can not follow the sinusoidal waveform anymore and tracks the signal with its maximum  $dV/dt$ .

quency. This constant value is called the *gain-bandwidth product* and with this value a designer can immediately determine the achieved bandwidth at a certain closed-loop amplification. Ideally the gain-bandwidth product is infinite.

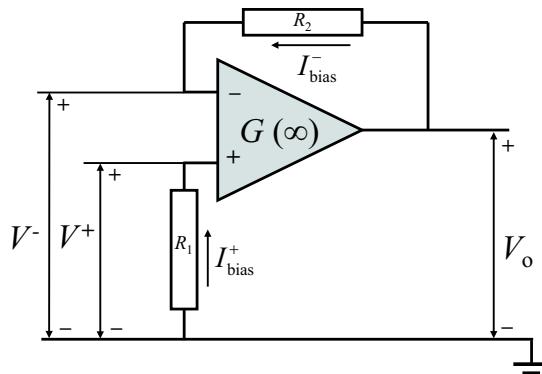
### Slew rate

Another consequence of the internal capacitances is the limited speed of change of the output voltage. This *slew-rate* is determined by the maximum current, that is available inside the operational amplifier to change the voltage over the internal capacitors, when amplifying high-frequency AC signals. The source of this current is limited by the nature of the design, which means that the voltage over the capacitors can not change faster, than a certain value of  $dV/dt$ . In practice this means, that at high frequencies the maximum amplitude of the signal is limited. Above these levels a sinusoidal waveform will change into a phase shifted triangular waveform with fixed slopes, because of the maximum  $dV/dt$ . Ideally the slew-rate is infinite.

#### 6.2.8.2 Limitations on the inputs

##### Input offset voltage and stability

The input *offset* voltage is the voltage needed between the + and - inputs, to achieve 0 V at the output. This is caused by small differences between the transistors in the differential gain stage of the operational amplifier. The presence of an offset voltage means in a feedback circuit, that the first rule in reality is a bit different. It should have been said that the amplifier will



**Figure 6.60:** The bias current in both inputs of an operational amplifier causes a voltage drop over the source resistors  $R_1$  and  $R_2$ . When both resistors are equal the effect on the output voltage is zero.

do whatever it can to achieve a difference between the input voltages, that is equal to the offset voltage. The offset voltage is only important in case of amplifying very small DC voltages, because in practice this voltage is smaller than a few mV. In some operational amplifiers it can be adjusted to near zero Volt. The long-term stability of the offset voltage, also called *drift*, is important to know, when very small DC voltages need to be amplified accurately over a long period of time. This drift is mainly caused by the temperature sensitivity of the transistors.

Ideally the offset voltage is zero without any drift.

### Input bias and offset current

The input bias and offset current represent the currents that flow into or out of the input terminals and that are required to operate the amplifier. In electronics *biasing* refers to adding a voltage or current to the useful signal in order to make the system work. The presence of this current implies, that also the second rule about zero currents in the inputs is not completely true. In a bipolar transistor input stage, like with the NE5532, the bias current is necessary to drive the first transistor with a current level in the order of several nA. With special transistors like the FETs in the OPA627, the input current can be reduced to several pA. A low value of the bias current is important when the path from the inputs to ground has a high resistive value. This current would otherwise influence the output voltage to compensate the related voltage drop over the resistors. The influence of the

bias current can be reduced by designing the amplifier in such a way that the source (Thevenin) impedance of the circuit around the + input equals the impedance of the circuit around the – input. This is shown in Figure 6.60 where the output voltage is zero in case of an equal bias current on both inputs and when both resistors are equal:

$$V^+ = 0 - R_1 I_{\text{bias}}^+ \quad (6.109)$$

$$V_o = V^- + R_2 I_{\text{bias}}^-$$

$$V^+ = V^- \implies V_o = R_2 I_{\text{bias}}^- - R_1 I_{\text{bias}}^+ = 0$$

It is clear that an offset in the bias current will result in an offset voltage at the output even when the resistors  $R_1$  and  $R_2$  are equal.

It is especially important to be aware that a capacitor as is applied in the feedback path of an integrator, is not able to conduct DC-currents like the bias current.

Ideally the bias current is zero.

### Noise voltage and current

Due to temperature, stochastic signals are generated in resistances and all other components of the operational amplifier. This noise is represented as a voltage over the inputs and as a current into the input. The noise voltage is directly translated into a noise voltage to the output, determined by the feedback circuit and the noise currents are translated to the output voltage, through the source impedance of the circuits around the inputs. The noise can be represented by a cumulative value, but is mostly represented by a noise density value as a function of the frequency area.

Ideally the noise is zero.

### Input resistance

The input resistance is the value that corresponds with the part of the input current that is dependent on the input voltage so excluding the bias current. This property is due to the transistors and resistors around the input circuits. Differential-mode is the resistance between both input terminals and common-mode is the resistance of both inputs to ground. In case of a feedback system, the differential-mode resistance plays a very limited role as the voltage between both inputs is kept close to zero. In reality the differential input resistance is reduced with a factor equal to the loop gain. With an inverting amplifier also the common-mode resistance is of little

importance as both input voltages are kept at ground level. Only with a non-inverting configuration, the common-mode input resistance has influence on the input currents.

Ideally the input resistance is infinite.

### **Input voltage range**

In most operational amplifiers, the inputs are not allowed to exceed the power supply voltage level and normal operation is only guaranteed over an even smaller range. Often the inputs are protected against occasional higher voltages by diodes but these will also reduce the input resistance and for that reason, protective input diodes are not used in very critical designs. Exceeding these values results in clipping of the output voltage to the maximum level and it can even cause damage to the device.

Ideally the input voltage range is infinite.

### **Common-mode rejection ratio**

The purpose of an operational amplifier is to amplify only the voltage difference between the two inputs, the differential-mode voltage. This means that the output voltage should be completely independent of the common-mode voltage of the input terminals. The ratio between the gain of the differential-mode to the gain of the common-mode is called the *common-mode rejection ratio* (CMRR).

Ideally the CMRR is infinite.

#### **6.2.8.3 Power supply and output limitations**

##### **Power supply rejection ratio**

Noise and level variations such as ripple of the power supply does preferably not have influence on the output voltage. The *power supply rejection ratio* (PSRR) is defined as the ratio between a power supply voltage change and an equivalent input voltage giving the same effect on the output.

Ideally the PSRR is infinite.

##### **Open-loop output resistance**

This determines the change of the output voltage as function of the output current. In a feedback system this value is reduced by a factor equal to

the loop gain. Because the loop gain is frequency dependent, due to the main pole, also the effective closed-loop output impedance will be frequency dependent. As it increases with frequency, this can be modelled as a very small series inductor.

### Power consumption

The power consumption is determined by many factors. The first factor is related to the idle current that is needed in the transistors and these are modulated by the signal. The power consumption is related to the slew-rate, as a high slew-rate requires a high internal current to charge the capacitors. Further also the maximum output current determines the power consumption.

The power consumption should ideally be zero and for portable devices special operational amplifiers are developed, that show a very low idle current level below 1 mA.

### 6.2.9 Closing remarks on low-power electronics

It is demonstrated that low power analogue electronics can be applied in many different ways, that all deal with the manipulation of electric signals by amplification and filtering. The operational amplifier has developed over time into an extremely versatile building block, optimised for different applications. With the basic knowledge of this chapter, it is possible to select suitable components in an analogue electronic system without help of specialists, but still too many things have remained untouched and need attention to achieve a real solid design. A few of them are summarised below:

- **The properties of the power supply.**

Especially at high frequencies the power supply will not be a very good voltage source anymore. This requires the addition of ceramic capacitors between the power supply and ground, as close as possible to the operational amplifier, to short high-frequency interference to ground, combined with a small electrolytic capacitor to dampen any high-frequency oscillations by its internal parasitic resistance. Sometimes even local linear power supply stabilisation is required. A stabilised power supply is a special version of a power amplifier, that will be presented in the following chapter. For moderate power levels,

voltage stabilising circuits are available as IC, that can directly be mounted on a printed circuit board.

- **Mains supply issues.**

The power socket of the mains is not really a nice 230 V AC voltage source. Sometimes the signal hardly looks like a sine wave at all. The worst are spikes of several kV, due to circuit breakers at other places or even caused by a lightning stroke. The electronic system needs additional safety components to lead away this excessive voltage from the sensitive parts.

- **Precautions for Electromagnetic compatibility (EMC).**

Electric current runs in loops and magnetic fields induce voltages in these loops. These voltages are added to the signal and can cause errors. This requires a careful lay out of the wiring of an electronic circuit. This subject will return in Chapter 8 on measurement.

- **Precautions against Electrostatic discharge (ESD).**

Especially ICs with MOSFET inputs are sensitive for electrostatic charge build up, that can damage the thin layer between the Gate and the channel. Besides careful handling in a discharged environment, this also means that the electronic circuit needs protective components to prevent electrostatic charge build up.

- **The effect of safety grounding.**

Most professional electronic equipment is provided with a grounded housing for human safety. This is not always the most preferred situation from an interference point of view. Also this subject will return in Chapter 8.

These and several other issues make the electronic field to an important, exciting and rather difficult specialisation, where a vast amount of knowledge is gained in the last decades.

## 6.3 Power amplifiers

Combined with the actuators and the mechanics, the power amplifier determines the dynamic behaviour of the mechatronic plant. Power amplifiers differ from the amplifiers in the previous section in their focus on reliable electrical power delivery and efficiency. As will be demonstrated, this focus has invoked several special technologies, that are only partly or hardly required in low power signal amplification. With the example of a modern tube amplifier according to ancient technology as shown in Figure 6.61 it is already clear that power is equal to size and heat but even with more efficient semiconductor devices like the ones shown in Figure 6.62 a power amplifier will remain always recognisable by the large components that have to sustain elevated voltages, currents and heat. It is also true to say that the design of power amplifiers will appear to be quite similar in certain ways, in spite of these differences. Especially the presented topology of the operational amplifier can be directly used for the design of linear power



**Figure 6.61:** Only 50 years ago proportional electronic control of high power could only be achieved by means of electron tubes working in vacuum at very high voltages with glowing wires to create free electrons. The shown audio power amplifier can deliver two times 80 Watt while producing about the same amount of heat. Fortunately semiconductors can do the job with less effort although also with less beauty.



**Figure 6.62:** The high power semiconductors as shown on the right can be distinguished from signal transistors as shown left by their mechanical means to connect them with a heat sink in order to keep the temperature of the semiconductor material below the level where the semiconductor stops working ( $\approx 150 - 200^\circ$ ).

amplifiers.

This section starts with an overview of the requirements of power amplifiers with an emphasis on the application with electromagnetic actuators. This will be followed by a subsection on linear amplifiers for moderate power levels with both a voltage and a current source output. The last subsection will present pulse-width Modulated switching output stages, that are used to reduce the dissipation at high power levels.

### 6.3.1 General properties of power amplifiers

The following main specification elements determine the design of a power amplifier.

- Power delivery capability.
- Dynamic properties.
- Output impedance.
- Efficiency.
- Linearity.

## Power delivery capability

A large power implies automatically a high value of the current and voltage. As it appears to be difficult to integrate high voltage electronics, the power stages in these amplifiers are often built with active and passive discrete components. These large coils, resistors, capacitors and high power semiconductors generally produce a lot of heat. Silicon semiconductors can withstand temperatures on the chip of  $150^\circ - 200^\circ$ , but due to the heat transfer resistance of the packaging, mostly the temperatures at the outside of the device have to be kept at a more reduced temperature level.

These thermal issues automatically require efficient methods to remove the heat from the sensitive electronics. The mechanical design of signal amplification electronics with thin printed circuit boards is a rather two dimensional exercise but the lay-out of a power amplifier is a real three dimensional mechanical challenge. The thermal design with cooling plates and the use of large parts requires a design, that takes the mass and the environmental mechanical dynamics under serious consideration. Solder joints for instance are not capable of sustaining high mechanical loads and, while this is a frequent cause of malfunctioning electronics, this is just an example of the special skills needed when designing power amplifiers.

Another issue of the high power is the potential interference by magnetic and electric fields, the *Electro Magnetic Compatibility* (EMC), caused by rapidly changing high voltages and currents. Also this aspect requires a careful three dimensional lay-out with shielding plates and conscious routing of the wiring.

## Dynamic properties

When a power amplifier is used in a mechatronic system, it determines a part of the feedback loop. A signal delay caused by the amplifier adds one or more poles with their related phase lag to the open-loop transfer function and reduces the stability margins. Especially with a digitally controlled amplifier (sample time) this is an aspect to pay attention to. In an analogue amplifier this delay is only related to the frequency response. Even when the amplifier is fast enough under normal resistive loads, the complex load of an electromagnetic actuator ( $L, R, C$ ) will influence the internal stability of the amplifier. For that reason, power amplifiers for high performance mechatronics need to be designed specifically for a certain application.

## **Output impedance**

A low output impedance of an amplifier is required when damping in the system is necessary, as was presented in Section 5.4.1.3 of Chapter 5. This is for instance the case if the amplifier is used in combination with loudspeakers, where the amplitude at the lower resonance frequency needs to be limited.

A high output impedance is useful when the current and the force needs to be independent of the movement and the self-inductance of the actuator windings. This is mostly the case in servo-controlled positioning systems with Lorentz actuators. With a high output impedance there is no force change caused by a relative velocity of the mover and stator, resulting in a reduced transmissibility between these parts when compared with a low output impedance amplifier. In Section 5.4.1 of Chapter 5 it was further shown that the inherent compensation of the self-inductance by a high output impedance has a certain danger, because too often mechatronic designers tend to think that the self-inductance is not important anymore, because of the availability of powerful electronics. In a high speed positioning servo-system this perception can have as consequence that the amplifier has to deliver a significantly higher voltage to overcome the high values of  $L(dI/dt)$  than is required to deliver the power in the system, because these maximum values mostly occur at different moments in a periodic cycle due to their reactive  $90^\circ$  phase relation. In that case the maximum voltage times the maximum current of the amplifier needs to be far higher than its power capability. This requirement can have a significant impact in the total system cost and reliability and is elaborated more in Section 6.3.2.2.

## **Efficiency**

An amplifier transforms a supply voltage and current into an actuator voltage and current. When this transformation is done by means of a linear element like a transistor, the actuator current has to run through this transistor. This actuator current multiplied with the voltage difference between the power supply and the output determines the power that is dissipated inside the amplifier. In case of the above mentioned high self-inductance, this inherently leads to a high power dissipation. As an ultimate solution a switched-mode pulse-width modulated amplifier is necessary to reduce or overcome the dissipation.

## Linearity, freedom of distortion

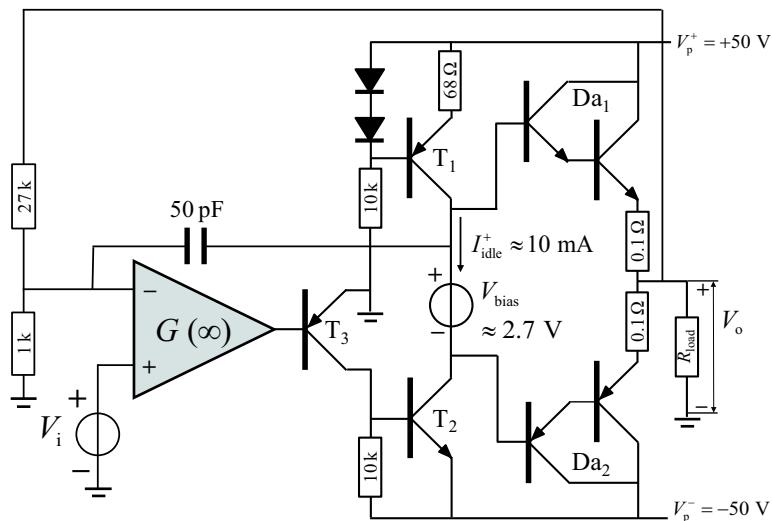
Non-linearity in an amplifier has several effects. In audio appliances it is the source of audible distortion, but in professional mechatronic systems the non-linearity impairs the behaviour in two other ways. First of all it changes the gain of the amplifier as function of the actual value of the output voltage or current. As a consequence the loop-gain in a feedback system becomes dependent on the actual output voltage with ultimately a chance to become unstable. In practice it is not difficult to limit the non-linearity to such a low level, that this instability issue hardly ever occurs, with one exception: The ultimate non-linearity happens when the output voltage has reached its maximum level, close to the supply voltage. At that moment the output voltage “clips” and all control is lost. This effect was introduced in Section 4.5 of Chapter 4, where it was explained how such extreme clipping can result in an uncontrolled “limit-cycling” of the motion system.

But even without this extreme example, a small non-linearity can cause errors with feed forward control. As long as the non-linearity is reproducible and deterministic it can be compensated, but non-linearity of an amplifier is a sign of a non-optimal design and that almost automatically implies the occurrence of thermal drift.

### 6.3.2 Linear power amplifiers

Figure 6.63 shows a basic design of a linear power amplifier with a voltage source output stage. At a first glance this design looks very familiar to a low power operational amplifier. There are however at least two significant differences. First of all the gain stage is designed to be able to deliver a high output voltage, significantly above the regular power supply voltage of  $\approx \pm 15$  V of the differential input stage. This is achieved with a *level-shifter* that consists of a discrete pair of high voltage transistors in voltage amplification mode. The second difference is the high-current power-output stage, that consists of a set of transistors with a very high current-amplification ratio, like the *Darlington pairs* that are named after the American electro technician Sidney Darlington (1906 – 1997), who invented the configuration. These pairs are in fact two cascaded emitter followers and can deliver currents up to several Ampères with a current-amplification of more than 1000. Also sometimes high-power MOSFETs are used as these do not require a continuous input current.

The function of the level-shifter is as follows. The input of transistor  $T_3$  is at a low voltage level and gets its input voltage from a low-power operational

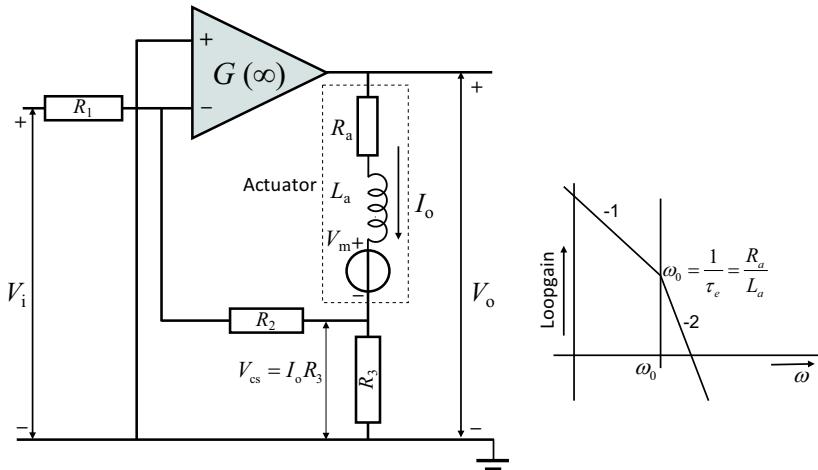


**Figure 6.63:** Basic design of a linear power amplifier with a voltage source output stage and a closed-loop gain of  $\approx 30$  dB, showing the main components. The transistors  $T_1$ ,  $T_2$  and  $T_3$  determine a level-shifter to achieve a large output voltage range. The push-pull output stage consists of two Darlington pairs  $Da_1$  and  $Da_2$ , that can handle a large output current up to 10 A. The operational amplifier serves as differential input and gain stage and the capacitor  $C$  creates the dominant pole for stability.

amplifier, acting as a combination of a differential input stage and a high-gain stage.  $T_3$  is a voltage amplifier with a collector resistor of  $10\text{ k}\Omega$ . The collector voltage of  $T_3$  is used as input for the base of  $T_2$ , that also acts as a voltage amplifier with a current source as collector resistor.

$T_1$  with its emitter resistor of  $60\text{ }\Omega$  and the two diodes, that give a constant base voltage for  $T_1$  define a current source with an idle current level of  $\approx 10\text{ mA}$ , taking the base-emitter threshold voltage of  $0.6\text{ V}$  of  $T_1$  into account. This configuration guarantees, that the voltage swing is not limited as with a resistor instead of a current source, the collector current of  $T_2$  would be dependent on the output voltage. Furthermore, a current level of  $10\text{ mA}$  is sufficient to drive the Darlington pairs up to an output current level of  $10\text{ A}$ , when the current-amplification of the Darlington pairs is  $\geq 1000$ .

The feedback circuit is arranged as a non-inverting amplifier with a gain of 28 which is equivalent to  $\approx 30$  dB. A separate feedback capacitor is added to create the dominant pole, while reducing the non-linearity of the high gain level-shifter. The resulting closed-loop bandwidth is  $100\text{ kHz}$ , determined by the RC-time of the capacitor with the feedback resistor of  $30\text{ k}\Omega$ .



**Figure 6.64:** Current-feedback is achieved by measuring the current with the series resistor  $R_3$  and using this voltage as feedback signal. Because of rule one the current through the actuator is only dependent on the input voltage and not on the output voltage. Unfortunately the self-inductance introduces an additional pole that can cause instability.

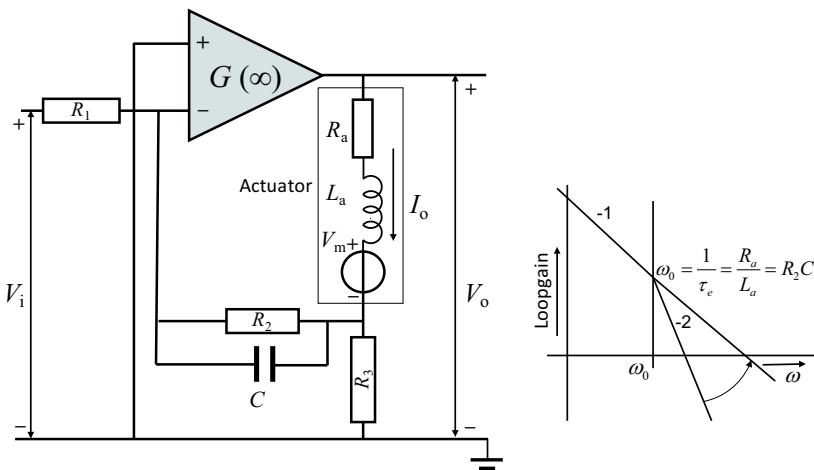
### 6.3.2.1 High output impedance amplifiers

The electromagnetic actuators that are used in most precision mechatronic motion systems require a power amplifier with a high output impedance. The previously presented high power output stage has a low output impedance, due to the emitter follower configuration and the voltage-feedback loop.

#### Current-feedback

This problem can be solved by creating a transconductance amplifier where the high output impedance is created by means of current- instead of voltage-feedback. This principle was introduced in the previous section and is repeated here in Figure 6.64 to explain the issues that are related to this configuration when used with a reactive load. In this case the power amplifier is symbolised as a standard operational amplifier and can be assumed identical to the amplifier of Figure 6.63 without the voltage-feedback circuit. The current is measured with a small series resistor  $R_3$  and the voltage over this resistor is fed back to the negative input of the power amplifier.

As was shown in Section 6.2.4.6 this configuration creates an output current that is only depending on the input voltage and independent of the actuator



**Figure 6.65:** Pole compensation by a parallel capacitor over the feedback resistor cancels the effect of the self-inductance but impairs the current measurement and reduces the closed-loop gain at higher frequencies. This can in principle be compensated by an additional differentiating lead-network in the controller.

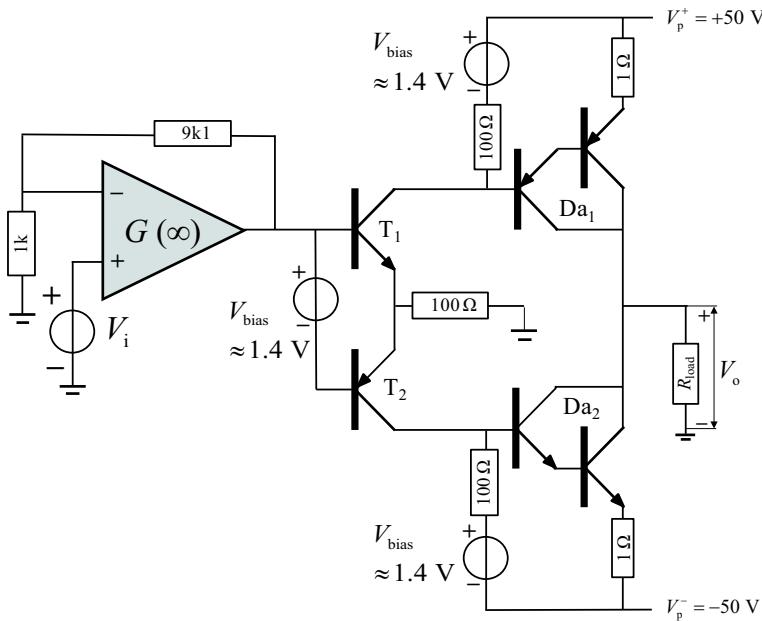
voltage. Unfortunately this is only true when the feedback loop has sufficient closed-loop gain and phase margin. With a real electromagnetic actuator the self-inductance of the actuator-windings causes problems, because above a certain frequency the impedance of the self-inductance will determine the impedance of the actuator as was explained in Section 5.4.1.1. This frequency equals:

$$\omega_e = \frac{1}{\tau_e} = \frac{R_a}{L_a} \quad (6.110)$$

Above this frequency, the feedback path from the output of the amplifier to  $V_{cs}$  will have a low pass characteristic with one pole and a  $-1$  slope that adds to the internal  $-1$  slope of the operational amplifier. This results in a marginal stability of the feedback system as is visible at the combined  $-2$  slope in the simplified amplitude Bode plot of Figure 6.64. This effect could be compensated by creating a differentiating action, a lead-network in the loop. This is possible by adding a parallel capacitor over  $R_2$  with the value:

$$\tau_e = R_2 C = \frac{L_a}{R_a} \quad (6.111)$$

This would indeed increase the phase margin as shown in Figure 6.65 but it would also change the closed-loop current source characteristics of the



**Figure 6.66:** By reversing the Darlington pairs of Figure 6.63 a current source output is obtained. This amplifier delivers 10 A output current at 1 V input voltage. For the positive half of the signal, transistor  $T_1$  acts as a level-shifting voltage amplifier with a gain of one,  $Da_1$  converts the voltage of the collector of  $T_1$  into an output current. The negative cycle is taken care of by  $T_2$  and  $Da_2$  with their surrounding resistors. The operational amplifier only acts as a non-inverting amplifier with a gain of 10, without feedback over the total amplifier.

system, that in its turn could be compensated by an additional lead-network in the controller. It is up to the reader to work this further out as an exercise.

Generally it is better to avoid these compensation circuits as they are never perfectly tuned. The solution is that a power output stage is designed with a high output-impedance in open-loop as shown in Figure 6.66.

### Current-source output stage

In this extreme example the current-feedback loop is completely avoided and the voltage amplifier level-shifter transistors and output stage are linearised with large emitter resistors. Furthermore the output stage of the amplifier is designed directly with a high impedance.

In this example, the operational amplifier acts only as a non-inverting

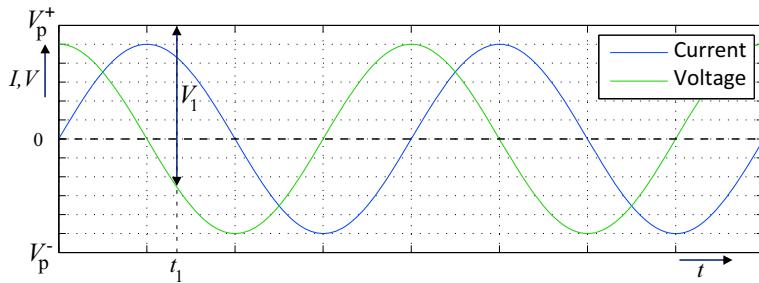
voltage amplifier with a gain of ten and a low output impedance and its output voltage is used as the input for the two level-shifter voltage amplifying transistors  $T_1$  and  $T_2$ . The current in these transistors is determined by the resistor of  $100\ \Omega$ , that is shared to linearise the transition between the positive and the negative cycle. In the positive cycle of the signal, the current in  $T_1$  will increase, resulting in an equal decrease of the voltage at the base of the Darlington pair  $Da_1$ , because of the  $100\ \Omega$  collector resistor. This results in an increased voltage over the  $1\ \Omega$  resistor in the emitter of  $Da_1$ , with a corresponding increase of the positive output current. For the negative cycle  $T_2$  and  $Da_2$  act the same.

A special aspect in this configuration is the equal value for the shared emitter resistor and the collector resistors of the level-shifters. This is done to linearise the amplification of the level-shifters, because there is no further feedback that could correct any non-linearity. This is also the reason for the relatively high value of the emitter resistors of the Darlington pairs. The low value of  $100\ \Omega$  around the level-shifters is chosen such that the current in the level-shifters at the maximum output current of  $10\ A$  is more than a factor ten higher than the input current of the Darlington pairs, again for linearisation reasons. At  $10\ A$  output current, the corresponding  $10\ V$  voltage over the  $100\ \Omega$  resistors requires a current in the level-shifters of  $100\ mA$ . The input current of the Darlington pairs is then  $10\ mA$  because of the current-amplification ratio and that is only  $10\ %$  of the current in the level-shifters.

It is clear from this example that all these measures come at a price in power dissipation and performance. The relatively high currents and resistor values both limit the output voltage and require significant measures to keep the temperatures of the semiconductors and the other parts at an acceptable level.

For that reason a more optimal solution will be to combine the current-feedback configuration with grounded load of Section 6.2.4.6 with the output stage from Figure 6.66, while using smaller emitter resistors to limit the power dissipation. The current-feedback will reduce any non-linearity and the collector follower configuration will reduce the problem of the additional pole by the self-inductance of the actuator.

At the end of this chapter, when presenting three-phase resonant-mode amplifiers, an alternative loss-less current sensing method will be shown which gives an almost ideal amplifier configuration and can be used also with linear power amplifiers.



**Figure 6.67:** With a complex load, the current from the amplifier will have a phase difference with the voltage. This means, that the amplifier needs to have a four quadrant capability by delivering positive and negative currents at any sign of the output voltage. This requirement implies a very high power dissipation.

### 6.3.2.2 Dynamic loads, four-quadrant operation

When an inductive or capacitive load is applied, the power dissipation in a linear amplifier is even more severe, because in that situation the current is out of phase with the voltage. As was explained in Chapter 5, electromagnetic actuators in mechatronic positioning systems show a sometimes large self-inductance and with piezoelectric actuators the load is mainly capacitive. In the extreme case of a purely reactive load the maximum current needs to be delivered at zero voltage, while at a quarter of the period a positive current is delivered with a negative voltage and at another quarter it is just the other way around, like shown in Figure 6.67. The capability to deliver currents with a different sign from the voltage at any moment is called a *four quadrant* operation. In linear amplifiers, the requirement to drive complex loads implies that the voltage difference over the power transistors at certain current levels can become even larger than the supply voltage. For instance in Figure 6.67 at time  $t_1$  the still high positive current has to be delivered by the power transistor connected to the positive power supply, while the output voltage is still negative. The voltage over the transistor that conducts the current is then as large as  $V_1$ , which is more than  $V_p^+$ .

Fortunately, reality is not always this extreme, and when the resistive part of the impedance of the actuator is large, the negative effect is reduced. This is the case in most audio applications where the reactive currents in loudspeakers are quite limited.

### **Motion induced voltage**

In mechatronic positioning systems the real problem is caused by the motion voltage of the actuator when high velocities are applied. At acceleration, the motion voltage increases in phase with the current and electric power is inserted in the system and converted into kinetic energy. The deceleration phase is however completely the opposite. While the motion voltage still has the same sign as with the acceleration, the current needs to be reversed and the full amount of kinetic energy has to be absorbed by the amplifier. Especially for this reason, the use of linear amplifiers in mechatronic systems is restricted to moderately low levels of power, with limited velocities, like in short stroke Lorentz actuators.

Because of these thermal issues and an increasing need for higher output power, the *switched-mode power amplifiers* are developed, that base their operation principle on alternately exchanging energy in inductors and capacitors.

### 6.3.3 Switched-mode power amplifiers

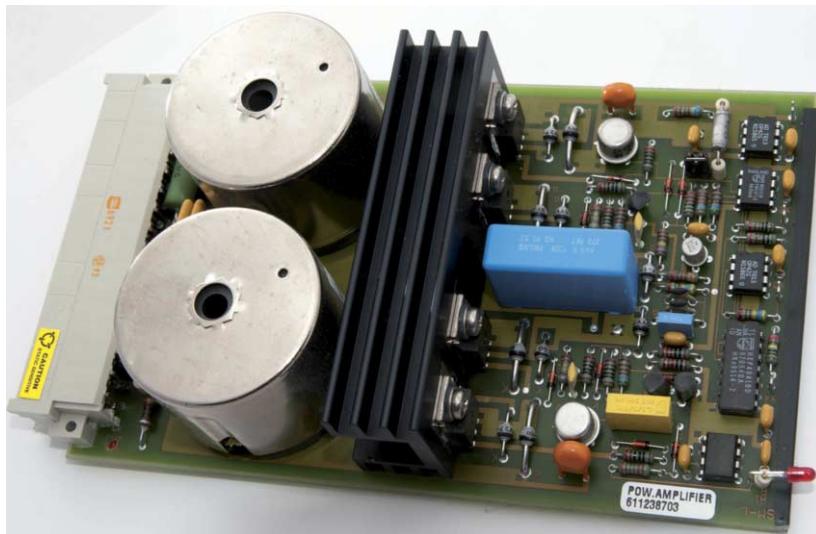
At the start of Chapter 6, the capacitor and inductor were introduced as reactive components, capable to store electric energy. It was also shown that filters can be designed with capacitors and inductors that have resonating properties. At resonance the energy flows alternately between the capacitor and the inductor. Furthermore a voltage over a capacitor can not be changed infinitely fast, but needs a current to change. The same is true for the current in an inductor, that needs a voltage to change. These effects are applied in the following section where electronic switches are used to direct the currents and voltages in reactive elements in order to create an amplifier.

Two typical switched-mode amplifiers will get an illustrating role in this section, both applied in the wafer stages of ASML. The first example as shown in Figure 6.68 is an early prototype, designed around 1980 at Philips “NatLab” and was used to drive the first electric wafer stage in the Silicon Repeater as was presented in Chapter 1. This amplifier delivered moderate output voltages with  $\pm 30$  V and currents up to 5 A. The second amplifier is a three-phase high-power amplifier, designed around 1995 at the Philips Centre For Technology and its basic principle is still used to drive the 5 kW long-stroke actuators in the Twinscan wafer scanners.

After an introduction of the first example amplifier, this section continues with the required fast electronic switches with the power MOSFET as the main element. This is followed by an explanation of the principle of pulse-width modulation in a switched-mode voltage-source amplifier with the first amplifier as example. With this basic understanding, a deeper analysis of current-flows in switching power-output stages is presented, finalised by an optimal design method, using the resonant-mode of the reactive components in the three-phase ultra-high power amplifier of the second example.

#### 6.3.3.1 First example amplifier

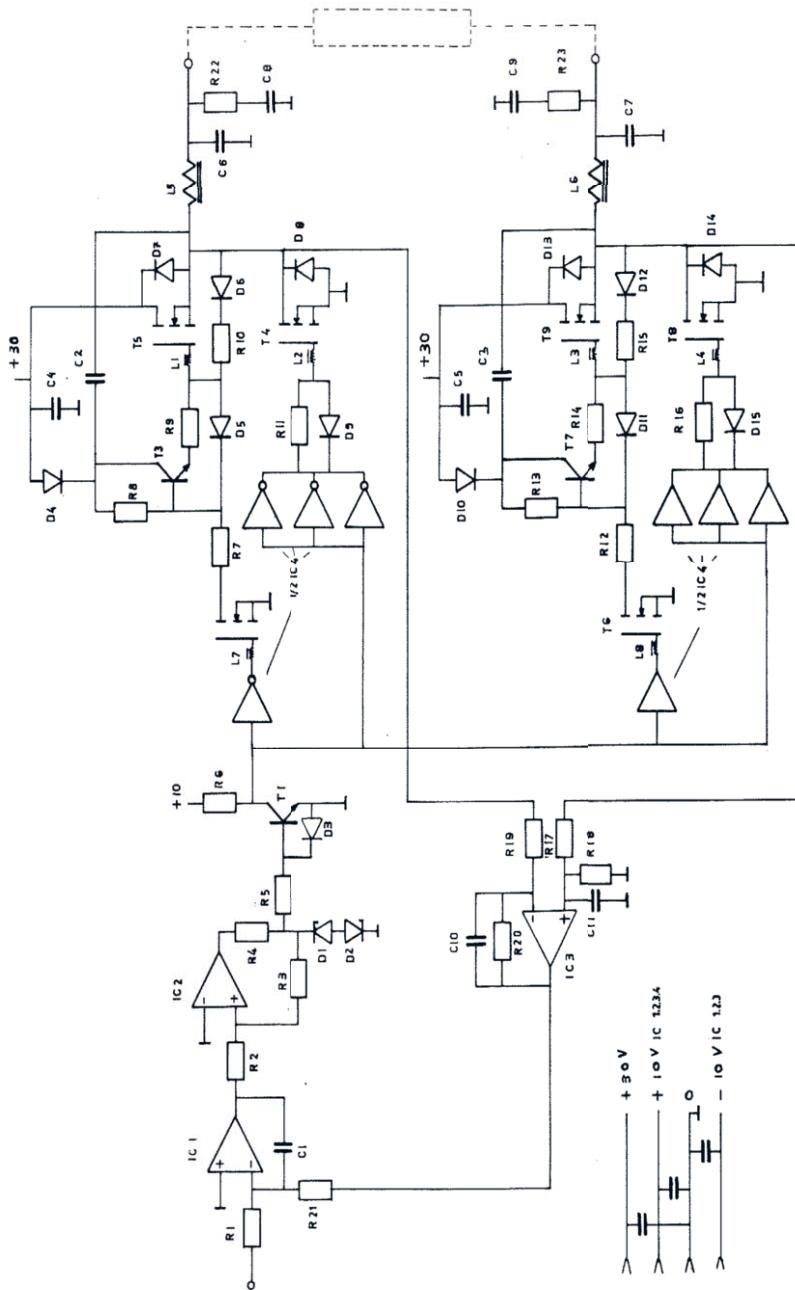
The first example amplifier is a proven design of a pulse-width modulated switched-mode amplifier. A copy of the original hand drawing of the circuit diagram from 1980 is shown in Figure 6.69. This amplifier was designed for driving the linear electric motors of the wafer stage, shown on the cover of this book for the Silicon Repeater Mark 2. The circuit diagram gives an indication of the relative simplicity of the design, with some digital switching components, like C-MOS integrated circuits (the little triangles), two operational amplifiers, six power MOSFETs, three bipolar transistors and



**Figure 6.68:** The switched-mode power amplifier from Figure 6.69 that was designed for the first electric wafer stage for the Silicon Repeater wafer stepper.

several passive components. Because of its straightforward functionality, this amplifier is used as example when explaining the working principle of switched-mode amplifiers and the functionality of every main building block will be illustrated by using the actual component values and other system properties, that will become clear in this chapter.

- The switching frequency  $f_s$  of this amplifier is 150 kHz at 0 V output voltage.
- The switching output stage consists of two single-ended switching power units in counter phase using only a single supply voltage of 30 V. This is equivalent to one single-ended output stage with two supply voltages of plus and minus 30 V.
- The output filter consists of an inductor with a self-inductance of 1 mH and a capacitor of 0.33  $\mu$ F, giving a corner-frequency  $f_0$  of 8.7 kHz with a  $-2$  slope in the attenuation band.
- The impedance of the actuator consists of a resistor of 5  $\Omega$  in series with a self-inductance of 8 mH, giving an electric time constant  $\tau_e = L/R = 1.6$  ms



**Figure 6.69:** An early example of a pulse-width modulated switched-mode power amplifier, designed for the linear motors of the first electrical wafer stepper of Philips Electronics and ASML. (Original hand drawing of the circuit diagram from 1980.)

### 6.3.3.2 Power MOSFET, a fast high-power switch

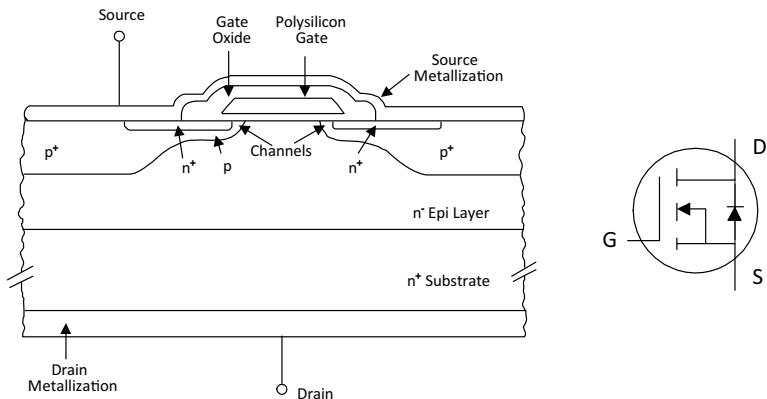
The currents and voltages in switched-mode amplifiers need to be switched at a very high frequency in the order of 100 kHz or more. An ideal electronic switch is able to conduct current in both directions, has zero resistance at closure, can withstand very high currents and voltages and has zero switching time. And last but not least an ideal electronic switch does not require energy to change its state.

It is clear that also in this case “ideal” does not exist, although the continuous improvements in electronic technology have brought us near ideal devices named power MOSFETs.

In precision mechatronic positioning systems, bipolar transistors are seldom used as switching devices. The reason is, that they require a relatively high base current. Further they are relatively slow, due to the low speed of the holes as charge carrier and they suffer from a phenomenon called *secondary breakdown*, a destructive mode that limits their power capability. On the other hand their ruggedness has made them popular in power-conversion systems, that do not need a very high switching frequency. The high base current has been solved by integrating a MOSFET as input device before the bipolar transistor. This combined device is called an *Insulated Gate Bipolar Transistor* (IGBT) and can switch several kilo Volt with several kilo Ampère current. Because of their relatively high switching times in the order of several tens of microseconds they are not preferred in high precision mechatronic positioning systems but found their application in electrical transport systems like trains and hybrid cars.

For the more high-frequency applications always N-channel power MOSFETs are used, as they offer switching speeds in the order of a few nanoseconds, because they use only electrons as charge carriers. Power MOSFETs are created by placing millions of small MOSFETs in parallel. This is allowed as each separate MOSFET behaves like a variable resistor and resistors can be placed parallel to create a smaller resistor value. This parallelism requires a special configuration of the MOSFET elements, different from the horizontal configuration that was shown in Figure 6.34. The vertical structure from Figure 6.70, with the Drain at the bottom side of the chip and the Source at the top side, enables to directly connect all Drains and all Sources, while the Gates are connected via an insulated embedded wiring network spread over the chip.

One special property of a MOSFET has not been mentioned yet. Most certainly the reader will have recognised, that a MOSFET is an essential



**Figure 6.70:** A power MOSFET consists of millions of small MOSFETs connected in parallel. This is accomplished by a vertical design where the Source and the Drain are at two sides of the chip and the Gates are embedded in an insulating oxide layer. The symbol shows that the substrate is connected to the Source. As a direct consequence of this connection, an additional diode is present between the Drain and the Source, determined by the junction between the Drain and the substrate. This diode is used effectively in switched-mode amplifiers. (Courtesy of International Rectifier)

symmetric structure, where the current can flow in both directions. Indeed that is the case and although these devices are optimised to work with currents running in only one direction, especially in power MOSFETs this phenomenon is used to advantage in switched-mode amplifiers. This optimisation to work in only one direction has a direct connection with the practice to connect the Source directly to the substrate, as they will have the same potential, when the current runs from the Drain to the Source in an N-channel MOSFET. As a consequence, the junction between the N-material of the Drain and the P-material of the substrate determines a Silicon diode between the Source and the Drain. This diode helps conducting current from the Source to the Drain, opposite to the normal current direction and it will be demonstrated, that this is a very useful property in switching inductive loads.

Like all other elements in electronics, also power MOSFETs suffer from several parasitic capacitances and for that reason they are not really ideal. The most important is the capacitance of the Gate, that amounts in the order of 10 nF for a power MOSFET that can switch more than 200 A. Also the capacitance between the Drain and the Gate plays a similar negative

role as the collector-base capacitance of a bipolar transistor. Its Miller capacitance as experienced at the Gate is fortunately less than the input capacitance, but still the rather large total capacitance requires a charging and discharging current each time the MOSFET switch needs to change its state. As a result a net current is flowing in and out of the Gate. In high power applications the switching voltage at the Gate needs to be at least 10 V and with a switching frequency  $f_s$  of 100 kHz, an input capacitance of 10 nF gives an effective current  $I_g$  to the Gate of:

$$I_g = f_s \frac{1}{2} C V^2 = 10^5 \cdot 10^{-8} \cdot 100 = 0.1 \quad [\text{A}] \quad (6.112)$$

This value needs to be compared with the maximum current of 200 A that can be switched. It is still below the value of a comparable Darlington pair in a continuous current situation, with a current-amplification factor  $\beta = 1000$ , and this is without the additional base current necessary to charge the Miller capacitance. In fact this current level to the Gate is quite acceptable. It should be noted however, that this current is directly proportional to the switching speed. It is one of the sources of power dissipation in a switched-mode power amplifier.

### 6.3.3.3 Pulse-width modulation

switched-mode power amplifiers operate according to the principle of pulse-width modulation. This principle is based on the understanding from Fourier analysis, that a periodic signal can be decomposed into one constant value  $a_0$ , representing the average DC value of the signal and a set of higher harmonics, of which the lowest frequency equals the temporal frequency of the periodic signal. By filtering these harmonics with a low-pass filter, only the average value remains.

For a square-wave signal the average value is zero, when the negative cycle is equal in time to the positive cycle. The ratio between the positive cycle and the full period is called the *duty-cycle*. Because of this definition, the average DC value  $V_{DC}$  of a square-wave signal is equal to the following expression, where  $V^-$  is the negative voltage of the square-wave signal and  $V^+$  the positive value:

$$V_{DC} = V^- + (V^+ - V^-) \cdot \text{duty - cycle.} \quad (6.113)$$

With  $V^+ = -V^-$  a duty-cycle of 50 % gives an average voltage of zero Volt.

By modulating the duty-cycle of a square-wave signal with a constant amplitude, the average value of the square-wave signal can be changed. This

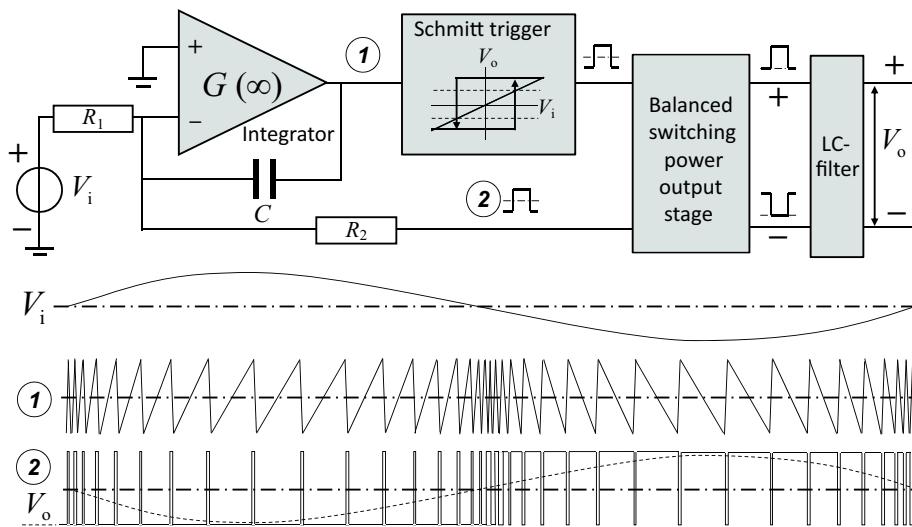
process is called *pulse-width modulation*. When the square-wave signal is created by means of switches, alternately connecting the output of an amplifier to either a positive or an equal negative high power supply voltage, the average value of the output voltage is determined only by the duty-cycle, independent of a load connected to this output. This average voltage can be obtained by filtering the higher Fourier terms out of the signal with a passive low-pass filter, consisting of an inductor and a capacitor.

### Pulse-width modulator

The first step to create a switched-mode amplifier is to convert a low-frequency signal into a square-wave signal with an average value that is equal to the momentary value of the low-frequency signal. In the example amplifier this process is based on a triangle – square-wave pulse-width modulator and is explained with the help of Figure 6.71.

The heart of the modulator consists of an inverting integrator and a *Schmitt trigger*, named after the American scientist Otto Herbert Schmitt (1913 – 1998) who invented it. The output of the Schmitt trigger can only have two states. In this amplifier one state is a positive voltage and the other state is a negative voltage, both with an equal magnitude. The output can only change its state from negative to positive, when the input surpasses a certain positive threshold and it can change its state from positive to negative only when the input comes below a certain negative threshold. In this amplifier also both threshold levels are equal. The output of the comparator is used as input for the high power switching stage and the high-frequency part of the output of that stage is filtered out by a passive LC-filter, transferring only the average DC value of the output signal. The switching output signal before the filter is fed back to the integrator. First the principle will be explained when assuming that both outputs are equal, because this switched output is an almost equal square-waveform as the output of the comparator.

Starting with the comparator in the positive output state, the output of the inverting integrator will show a negative slope over time at a speed depending on the voltage of the comparator, the input voltage, the capacitor value and the resistor values of  $R_1$  and  $R_2$ . As soon as the output of the integrator has reached the negative threshold of the Schmitt trigger, the latter will switch its state to a negative voltage and the output of the integrator will start to rise again, until the positive threshold is reached. At that moment the Schmitt trigger changes its state back to the positive



**Figure 6.71:** pulse-width modulator as used in the switched-mode power amplifier of Figure 6.69. The sum of  $V_i$  and the output of the balanced switching output stage is integrated by the inverting integrator. In combination with the Schmitt trigger in a closed feedback loop, this creates an oscillator with an average output voltage proportional to the input voltage.

output voltage. This process continues, resulting in a square-wave signal at the output of the Schmitt trigger with a corresponding triangle wave signal at the output of the integrator. When the input voltage equals zero, the resulting square-wave and triangle wave will be fully symmetrical with both an average value of zero Volt. As soon as a positive input voltage is applied, the resulting current in  $R_1$  in the direction of the minus input of the integrator will add to the current in  $R_2$ . As a result the down slope of the integrator will become steeper relative to the up slope, resulting in a tilted triangle wave at the output of the integrator with a non-symmetrical square-wave at the output of the Schmitt trigger. The triangle wave output of the integrator will always remain zero in average as the gain of the integrator is infinite for zero Hertz. This means that the average value of its inputs also needs to be zero to fulfil rule one of the operational amplifier. This implies that the average value of the square-wave output of the Schmitt trigger is proportional to the input voltage in a ratio that is determined by  $R_1$  and  $R_2$ . By filtering the square-wave by means of the LC-filter, this average value is obtained at the output and the system acts like a normal inverting amplifier.

In reality the balanced switching output stage creates a high power version of the square-wave from the output of the Schmitt trigger. By measuring the output before the filtering, a signal is obtained that in case of an ideal power stage would be proportional to the output of the Schmitt trigger. Because nothing is ideal, this feedback of the real switched output signal will help to correct errors in this stage because of the high gain of the integrator. The Schmitt trigger in the example amplifier of Figure 6.69 is realised with positive feedback around the second operational amplifier (IC2).

### 6.3.3.4 High-power output stage

To investigate the relation between current, voltage and power, Figure 6.72 gives a detailed insight in the behaviour of a switching power output stage, including the reactive components.

To illustrate the effects with real numbers this example shows the signals with the actual component values of the example amplifier, a 150 kHz switching frequency with an output filter consisting of an inductance of 1 mH and a capacitor of 0.33 µF. An equivalent *single-ended* output stage is chosen with equal positive and negative supply voltages of 30 V. A duty-cycle of 50 % results in an average output voltage of 0 V.

One cycle of 6.66 µs is shown in the figure, equally divided in four steps.

The starting point of the first step is taken when the current in the inductor is zero and the first power MOSFET M1 is just switched on, transferring the full positive power supply voltage of 30 V to the voltage  $V_b$  at the *bridge* between the switches. As will become clear after the last step, the voltage  $V_o$  at the output of the filter is then at its maximum negative value of  $\approx -0.1V$  and never grows bigger than an alternating value of this same magnitude. Because of this low value of the output voltage at this duty-cycle, the total voltage difference over the inductor is only determined by the bridge-voltage  $V_b = 30$  V and as a result the current in the inductor of 1 mH will increase in the first 1.66 µs to a maximum level  $\hat{I}_L$  determined by:

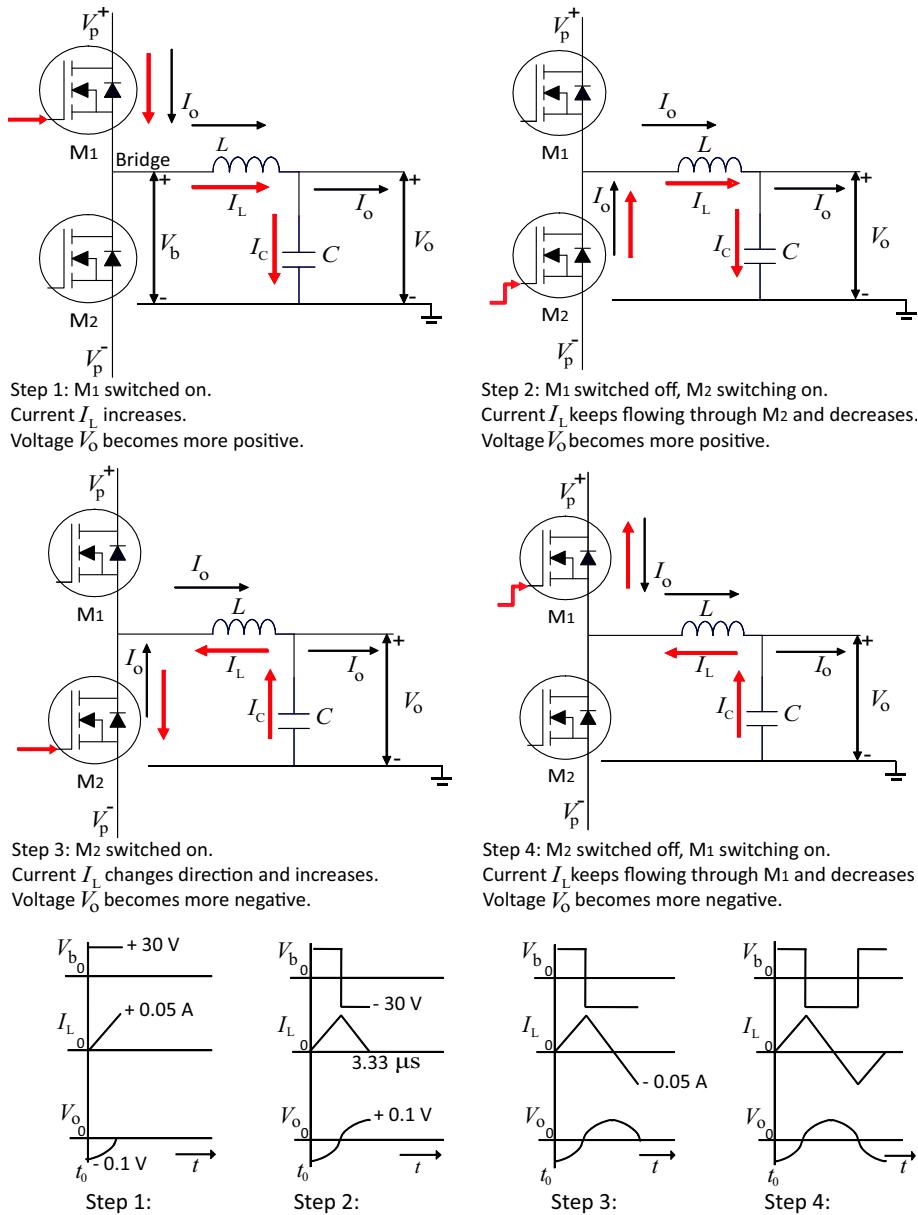
$$V_b = L \frac{dI}{dt} \quad \Rightarrow \quad \hat{I}_L = \int_0^{\tau} \frac{V_b(t)}{L} dt = \frac{V_b}{L} \tau \approx 0.05 \quad [\text{A}] \quad (6.114)$$

This current flows in the capacitor of 0.33 µF and the voltage change is calculated by integrating the current over the 1.66 µs:

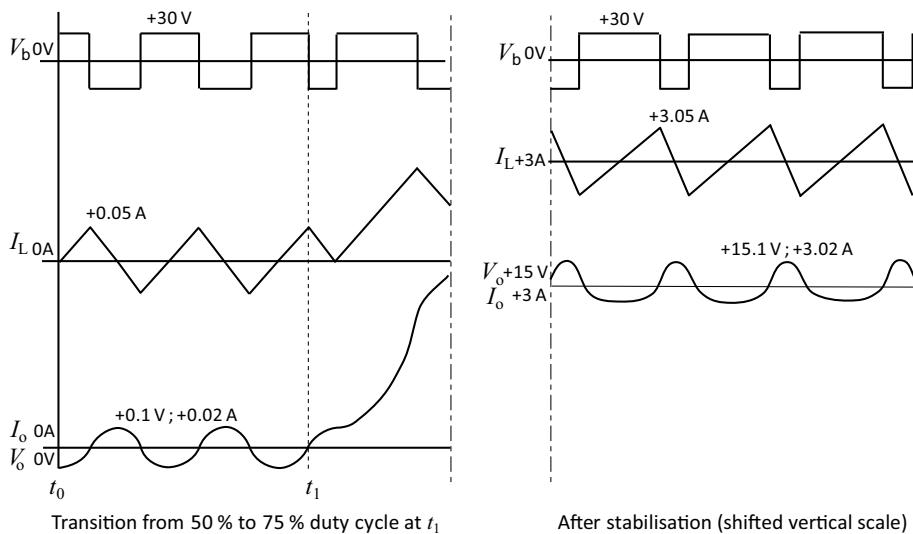
$$I_L = C \frac{dV_o}{dt} \quad \Rightarrow \quad \delta V_o = \int_0^{\tau} \frac{I_L(t)}{C} dt = 0.5 \frac{\hat{I}_L}{C} \tau \approx 0.1 \quad [\text{V}] \quad (6.115)$$

As a result the voltage  $V_o$  is zero after the first step.

In the second step M1 is switched off and M2 is switched on shortly after but not immediately, as that could cause a short-circuit current as will be explained later. However, in spite of this short delay, the voltage  $V_b$  will quite rapidly jump to the negative power supply voltage of -30 V after M1 is switched off. This is caused by the current in the inductor, that needs a voltage to change and will remain running until a negative voltage has reduced the current level to zero again. Because of the delay in switching of



**Figure 6.72:** Current flow in a switching power stage of a switched-mode power amplifier with power MOSFET switches and an LC-filter. The red arrows correspond with the momentary direction of the high-frequency current within a cycle and the black arrows correspond with the direction of a positive low-frequency current. One cycle consists of four equal steps of 1.66  $\mu$ s for each step, corresponding with an average output voltage of 0 V.



**Figure 6.73:** When the amplifier is loaded with a resistive load of  $5\ \Omega$  a change in the duty-cycle from 50 % to 75 % will cause an average output voltage of +15 V with a corresponding average output current of 3 A. At the transition moment  $t_0$  first the current and the voltage will increase until an equilibrium is reached. The ripple remains almost constant with only a change of waveform.

$M_2$ , the current will first flow through the diode of  $M_2$  before  $M_2$  is switched on completely. This proves the usefulness of this diode as without the diode the voltage  $V_b$  could be forced to more negative values than  $V_p^-$ , leading to damage of the MOSFET.

As a result of the negative voltage over  $L$  the current  $I_L$  will decrease again until it is zero at the end of the second step.

The third and the fourth step are identical to the first and second step. Only the signs are different and as a result the end of the fourth step is identical to the beginning of the first step.

In this still quite ideal situation it is clear than no power is dissipated in any element. The switching MOSFETs either conduct current without a voltage between their terminals and the reactive elements only store energy.

The reasoning from Figure 6.72 is true for the situation with 50 % duty-cycle and zero average output voltage. In that case also the average output current  $I_o$  to a resistive load is zero. This amplifier was designed for a  $5\ \Omega$  load and Figure 6.73 shows what happens with the current, when the duty-cycle is changed to 75 % and the load of  $5\ \Omega$  is applied. At that duty-cycle the

average output voltage will be 15 V and the corresponding average output current is 3 A. As soon as the duty-cycle is changed, the current  $I_L$  in the inductor  $L_c$  will increase, with a corresponding more continuous increase of the output voltage  $V_o$  and output current  $I_o$ . A higher level of  $V_o$  has as consequence that the voltage difference over the inductor at the upward current slope is reduced and the voltage difference at the downward current slope is increased. This causes these slopes to change until at equilibrium the slopes just correspond with the difference in timing. For instance at 75 % duty-cycle, the voltage difference in the upward current slope in the inductor is 15 V and the voltage difference at the downward current-slope is 45 V, while also the timing is 1 : 3. This means that the peak to peak AC level of the current is unchanged and this also means that the peak-to-peak level of the voltage is not changed. In fact only the shape of the signals is changed.

### Magnitude of the ripple voltage

The calculated ripple voltage of Equation (6.115) can be verified by using the attenuation of the LC-filter on the original square-wave signal. Because of the  $-2$  slope in the attenuation band of this second-order filter, the peak ripple voltage  $\hat{V}_{\text{ripple}}$  due to the peak output voltage  $\hat{V}_b$  at the switching output equals approximately:

$$\hat{V}_{\text{ripple}} = \hat{V}_b \left( \frac{f_0}{f_s} \right)^2 = 30 \left( \frac{8.6}{150} \right)^2 \approx 100 \quad [\text{mV}] \quad (6.116)$$

This corresponds with the value found in the previous analysis and is about  $-50$  dB below the maximum voltage level of the switching output. It is important to note that this ripple voltage is always present, even without a low-frequency output signal. The only way to reduce this ripple would be by either increasing the switching frequency, by taking a higher order filter, or by lowering the corner-frequency of the filter. In one way this illustrates the need for very fast switches when these amplifiers are applied in high precision positioning systems. On the other hand a signal of 100 mV at 150 kHz is not expected to give much acceleration in a heavy positioning system and it is also not audible too.

Like with all decisions in mechatronic design, for every specific situation it has to be examined, whether these artefacts would give a problem or not.

### 6.3.3.5 Preliminary conclusions and other issues

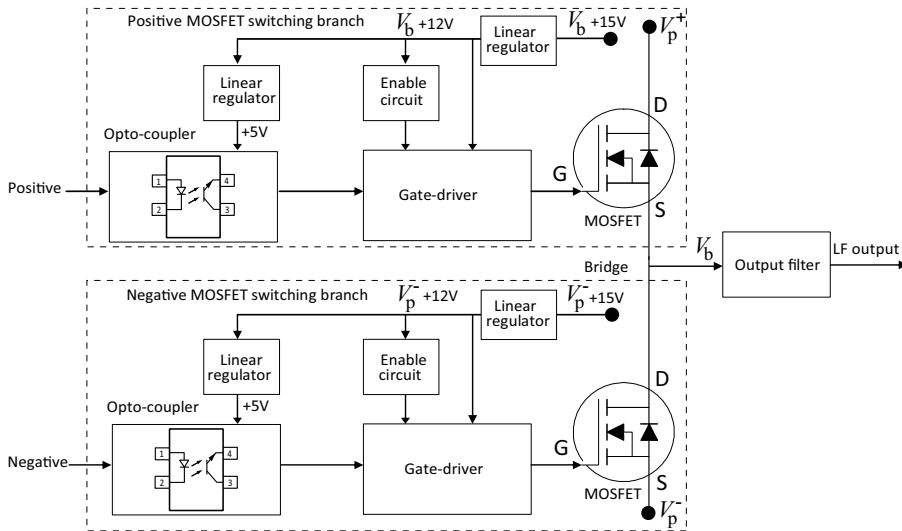
With these results it is shown that in principle an amplifier can be created that gives a low-frequency output signal proportional to the input signal by modulating a high-frequency switching signal and filtering the high-frequency components. No power dissipation takes place, because there are only switches and reactive components and all would be well when there were not some issues that still need attention.

- The Gates of the N-channel MOSFETs need to be driven relative to the Source.
- *charge-pumping* occurs with single-ended output stages.
- The interaction between the output filter and the load impedance.
- Speed limitations of the power MOSFETs.
- The need for a current source output impedance.
- The need for three-phase amplification

These issues are all addressed in the following subsections to complete this chapter on power amplifiers.

### 6.3.3.6 Driving the power MOSFETs

A specific requirement of power MOSFETs is that they need a voltage of around 10 V between the Gate and the Source to be switched on. Also the use of only N-channel MOSFETs poses special requirements on the electronic circuitry to drive the Gates, while the high input capacitance at the Gate requires the driver to be capable of delivering a very high current in an extremely short time. For the MOSFET at the negative supply side, this means that the switching voltage has to be supplied between the Source at the negative supply voltage  $V_p^-$  and the Gate. For the MOSFET at the positive supply side, the Source is even not at a constant voltage as the bridge-voltage  $V_b$  of the output switches from  $V_p^-$  to  $V_p^+$ . The controller electronics always operate at a low voltage (5 – 10 V) around ground level. As a consequence of these requirements a *floating* Gate-drive circuit needs to be used as shown in Figure 6.74. A floating Gate-drive circuit has its own power supply, galvanic insulated from the power supply of the amplifier and the switching commands are transferred by an insulating opto-coupler. A



**Figure 6.74:** Gate drivers for the switching power MOSFETs need to relate the Gate voltage to the Source voltage, that is either at the negative supply voltage  $V_p^-$  or at the bridge-voltage  $V_b$ . These voltages are different from ground and  $V_b$  is even fast switching between both supply voltages which means that the Gate driving circuits must be insulated from the control part. This is done by means of opto-couplers.

(Courtesy of ASML)

photo coupler consists of an LED with a photo-transistor, which is a bipolar transistor where the base is exposed to the light that has to be measured. The photons that fall on the exposed base will excite electrons that act like the regular base current in a normal transistor, resulting in a current at the collector when the transistor is supplied with a suitable voltage. One can also use a photo-diode for this purpose where the photons excite electrons in the depletion layer between the N- and P-dotted silicon but the sensitivity of a photo-transistor is higher because of the current amplification ratio which is often preferred.

The insulation of the power supply is not shown in the figure. In general this can be realised in the main power supply by using an electric transformer that is designed with a low capacitive coupling between the windings to prevent high-frequency currents by the switching bridge-voltage.

The linear regulators are devices that keep their output voltage at a constant level and reduce interference from other electronics that are connected to the same power supply.

### 6.3.3.7 Charge-pumping

The presented single configuration with two power supplies appears to give a problem with these energy sources, because the load is connected between its output and ground. The problem occurs when the amplifier delivers a low-frequency or DC current to the load and is explained with the help of the black arrows in Figure 6.72.

When a DC current is delivered to the load, this DC current is superimposed on the high-frequency AC currents as shown with the red arrows. When this positive current is larger than the amplitude of the high-frequency current, the total current in the inductor has become unidirectional in all steps of the switching process. For step one and four that is no problem because this positive current is supplied by the positive power supply. But in step two and three this positive current will be supplied from the negative power supply to ground. This means that the negative power supply delivers a negative power. Or in other words, the negative power supply receives power. This is also imaginable when recollecting the fact that the positive current equals a flow of negatively charged electrons and in step two and three these electrons move from ground towards the negative power supply. As a result of this process, the negative power supply gets more charge.

The energy involved in this process originally came from the positive power supply in step one where it was stored in the inductor and for that reason this process is called *charge-pumping*. With a positive DC output current of the amplifier, the negative supply should be capable to accept all this negative charge by delivering this positive current. Like with a linear power amplifier, most standard power supplies do not have this four-quadrant capability where both the positive and negative supply voltage can handle a positive and negative current. Take for instance the power supply that is made with the diode-bridge rectifier and storage capacitor from Figure 6.31 of the previous section. A positive current from the negative power supply, that is caused by charge-pumping, will result in a continuous increase of the negative supply voltage, because the diodes of the rectifier can not conduct current in the reverse direction. This increase of the negative supply voltage will continue until damage occurs in the storage capacitor or in other components.

The same effect but in the opposite direction happens with a negative DC current, in which case a continuous charge is transported from the negative to the positive power supply.

When only alternating currents are to be delivered to the load like in audio

power amplifiers, the storage capacitor can handle the short excess of charge in each signal period. Mechatronic systems often require a more extended period with a DC current component and in that case a solution for this charge-pumping phenomenon has to be found.

One solution can be to use a power supply with four-quadrant capability. In that case the resulting energy can for instance be dissipated somewhere inside the power supply in a resistor, although then a large part of the benefit of the switched-mode amplifier is lost. The best solution is to transfer the energy back to the mains supply. This is achieved in special switched-mode power supplies, that operate with the same principles as a switched-mode amplifier. By switching currents and voltages over reactive elements and the application of high-frequency transformers to change the voltage levels, they are optimised for DC voltages only, with as much as possible a voltage source output impedance. With additional switches the conversion from the mains power to the DC output power can be reversed but power supplies with that capability are more complex than power supplies that work in only one direction.

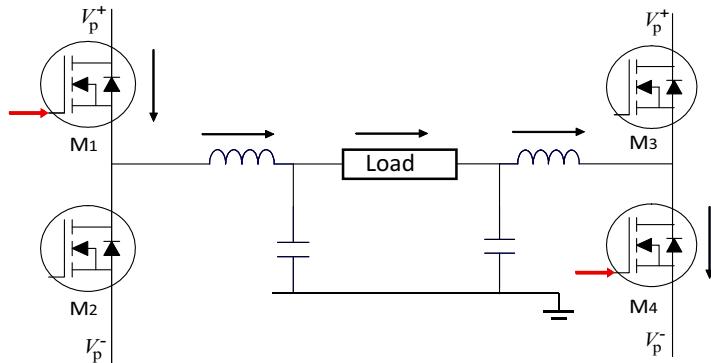
Another good example of a power supply with four-quadrant capability is a rechargeable battery, as long as the maximum charge level is not exceeded.

When the power supply is not able to handle these currents that run continuously in the opposite direction of the voltage, the problem can be solved in the amplifier itself, by using the *H-bridge* or *dual-ended* configuration as is applied in the example amplifier.

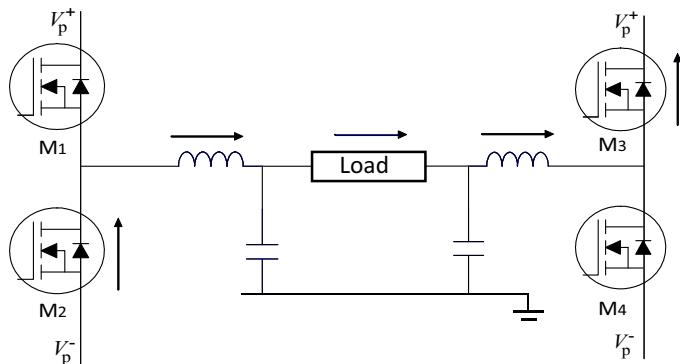
### 6.3.3.8 Dual-ended configuration

In the dual-ended H-bridge configuration, one terminal of the load is connected to one single-ended amplifier, while the other terminal is connected to a second single-ended amplifier. The second amplifier has to deliver the same voltage and current as the first amplifier but with a different sign. This means that both amplifiers share the same current at half of the required voltage.

One benefit of this configuration is the possibility to use a lower supply voltage or even a single supply voltage like in the example amplifier from Section 6.3.3.1. Using a single power supply reduces the complexity of the total system as one single-ended output stage is less complex than a complete second power supply. Human safety is also less affected because the maximum voltage difference at any location in the amplifier is only one time the power supply voltage. Furthermore, the possibility to ground the



Step 1: M<sub>1</sub> and M<sub>3</sub> switched on. The LF current flows from  $V_p^+$  to  $V_p^-$ .

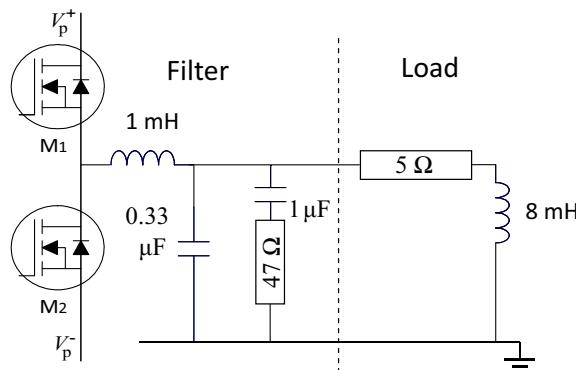


Step 2: M<sub>1</sub> and M<sub>3</sub> switched off. The LF current flows from  $V_p^-$  to  $V_p^+$ .

**Figure 6.75:** In a dual-ended H-bridge configuration the load is connected between two single-ended amplifiers that work in counter phase. The magnetic energy in the inductors, build up in step one is delivered by both power supplies and gained back in step two.

Source of the power MOSFET at the negative side avoids the need for an insulated Gate drive circuit for that power MOSFET.

The second advantage of a dual-ended configuration in a switched-mode amplifier is the absence of charge-pumping. In Figure 6.75 it is shown that, in the event of a positive DC current to the load at step one, the current flows from the positive supply voltage to the negative supply voltage. Both supplies are loaded with the same current and both deliver energy to the inductors and the load. At step 2 the current flows from the negative supply to the positive supply, while both supplies regain energy from the stored magnetic energy of the inductors.



**Figure 6.76:** An inductive load requires an additional RC network to correct the damping of the LC-filter in a switched-mode amplifier.

In this configuration each power supply has to absorb at maximum the magnetic energy that it inserted in the same switching cycle. This means that charge-pumping is fully avoided.

Because of these benefits the dual-ended configuration has become the de-facto standard in the industry for *single-phase* amplifiers, that drive only one load.

### 6.3.3.9 Output filter

When calculating the damping at the corner-frequency of the applied output filter in the example amplifier, one could have concluded that a 5 \$\Omega\$ load to the LC-filter with an inductor of 1 mH and a capacitor of 0.33 \$\mu\$F results in a quality factor of:

$$Q = R \sqrt{\frac{C}{L}} \approx 0.1 \quad (6.117)$$

This is a very low value and if this load was purely resistive in reality, a filter with an inductor of 0.1 mH and a capacitor of 3.3 \$\mu\$F would have been more appropriate. The load is however an electromagnetic actuator with a series inductance of 8 mH. This gives an electrical time constant of \$\tau\_e = 1.6\$ ms and as a result the actual impedance rises above \$\approx 100\$ Hz with a slope of +1 to become approximately 430 \$\Omega\$ around the corner-frequency of 8.6 kHz of the LC-filter. With that almost purely inductive value the LC-filter would have almost no damping anymore and noise on the input of the amplifier would be strongly amplified at that frequency.

To correct this side-effect of the complex load impedance, an additional RC-network needs to be added that determines the load impedance above approximately 1 kHz with a resistive value of  $47 \Omega$ , giving a  $Q$ -factor of almost one. In principle this resistor will dissipate power at higher frequencies, so it should not be chosen with a too low value. For that reason the inductor and capacitor values of the LC-filter were not chosen optimal for a resistor of  $5 \Omega$  but for this higher resistive value.

With a piezoelectric actuator this problem is quite different, as this kind of actuator gives a capacitive load, connected parallel to the capacitor of the LC-filter. This does not introduce additional stability problems but it also does not reduce the  $Q$ -factor. In that case, next to a parallel RC-network for damping, it might even be necessary to introduce a series inductor to the piezoelectric actuator in order to avoid high-frequency currents in the actuator.

From these short considerations it is clear that also the output filter section of a switched-mode amplifier is subject to dynamic optimisation in respect to the application.

### 6.3.4 Resonant-mode power amplifiers

The most important limitation of any component in a mechatronic system is determined by its dynamic properties. The power MOSFETs provide no exception to that rule. In this section a method will be presented, the *resonant-mode* output stage, that optimises the performance of a switched-mode output stage for these dynamic limitations by using the resonance of the inductor of the LC-filter with an additional capacitor.

#### Hard-switching

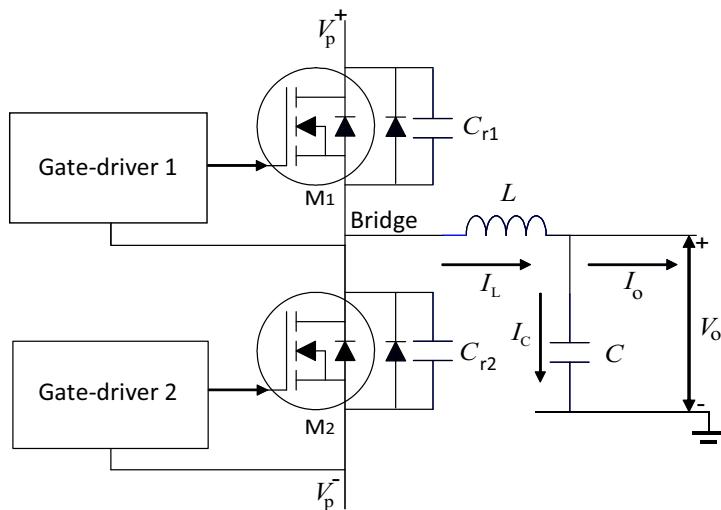
A major problem of the switched-mode power conversion as described above is the *hard-switching* transients that occur over the MOSFETs. At the moment that a MOSFET is switched off, the bridge-voltage changes very fast, forced by the current in the inductor that needs a voltage step to change. This voltage-transient causes a current through the parasitic capacitor between the Drain and the Gate of the switching MOSFET, that keeps the MOSFET conducting for a short time, resulting in a less steep voltage-transient. During this short moment, where the bridge-voltage is not equal to one of the supply voltages, the current through the MOSFET will cause power dissipation.

A second hard-switching effect is related to the conducting diode in step two and four, when the current flows in the opposite direction through the MOSFET. Even when the MOSFET is switched on, some current will still run through the diode. A semiconductor diode needs time to build up a new depletion layer to become non-conducting when the current reverses. This *reverse-recovery time* of the integrated diode in a MOSFET switching power transistor amounts up to approximately 100 – 500 ns. During this time a real short-circuit occurs, when the opposing power MOSFET is switched on. This problem can be alleviated by paying attention to three aspects:

1. Reduction of the current through the intrinsic diodes of the MOSFET. This can be achieved by adding very fast diodes with a low threshold voltage parallel to the power MOSFETs. For moderate power levels the *Schottky diodes* can serve that purpose. They are different from normal Silicon diodes as they are based on a combination of a metal and N-dotted silicon, to achieve a diode functionality. Unfortunately these diodes have a rather high leakage current in the reverse direction and the reverse voltage is limited to below 100 V. For that reason the use of Schottky diodes is limited to low voltage amplifiers. Fortunately also other special switching diodes have been developed for this purpose based on regular N- and P-material.
2. The MOSFETs should switch on with a little delay. Preferably the MOSFET is switched on, when its intrinsic diode is conducting, at least after the moment that the diode over the other MOSFET switch has stopped conducting.
3. The steepness of the transient slope of the bridge-voltage has to be reduced.

The second condition was already fulfilled in the 50 % duty-cycle situation of Figure 6.72 but as soon as the duty-cycle is changed to a value, where the current in the inductor becomes unidirectional, this condition is not fulfilled anymore.

The entire problem can be solved by tackling all items simultaneously and this is accomplished with a resonant version of a power output stage as shown in Figure 6.77. In this configuration three measures are taken. First the transient slope of the voltage is controlled by two additional capacitors, secondly the MOSFET is always switched on at zero current and thirdly fast parallel diodes are added to transfer the current in the reactive phase.



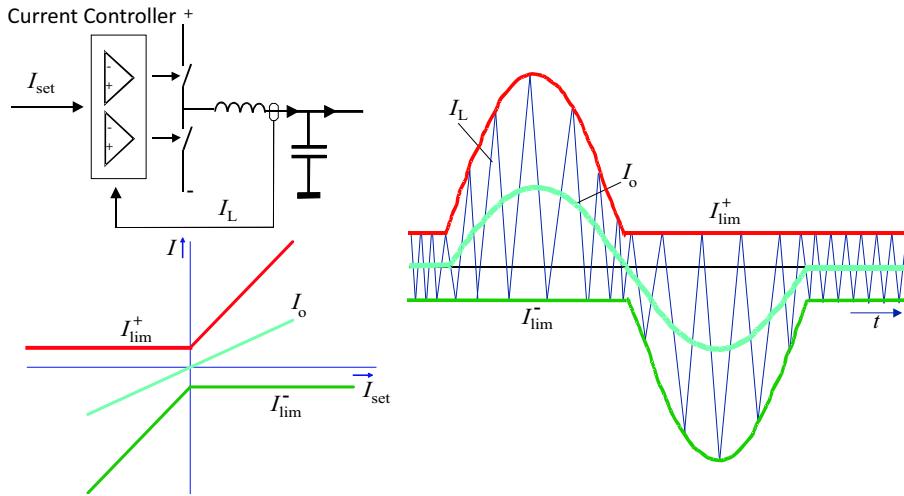
**Figure 6.77:** A resonant switching amplifier uses two additional capacitors  $C_{r1}$  and  $C_{r2}$  parallel to M1 and M2 that reduce the transient speed in a non-dissipative way to a sufficiently low level that the dissipation in the MOSFETs caused by their Drain-Gate capacitors is largely avoided. The additional parallel diodes are faster and have a slightly lower threshold voltage than the intrinsic diodes of the MOSFETs.

(Courtesy of ASML)

### 6.3.4.1 Switching sequence of the output stage

In Figure 6.78 the basic principle of the control of the resonant-mode output stage is shown for a single-ended output stage with two supply voltages. Because the problem originates in the current, this controller is essentially a current-feedback system. As a direct consequence of this controller, the amplifier becomes a current source amplifier, like is preferred in a mechatronic positioning system.

The controller continuously monitors the current  $I_L$  through the inductor and switches the MOSFETs when  $I_L$  has reached certain predetermined reference levels. These levels are always positive for the MOSFET that switches the positive voltage and negative for the MOSFET that switches the negative voltage. In case of a positive output current, the positive reference level is increased and with a negative output current, the negative reference level is increased. The MOSFETs are switched on again as soon as their current is reversed at the moment of zero current.

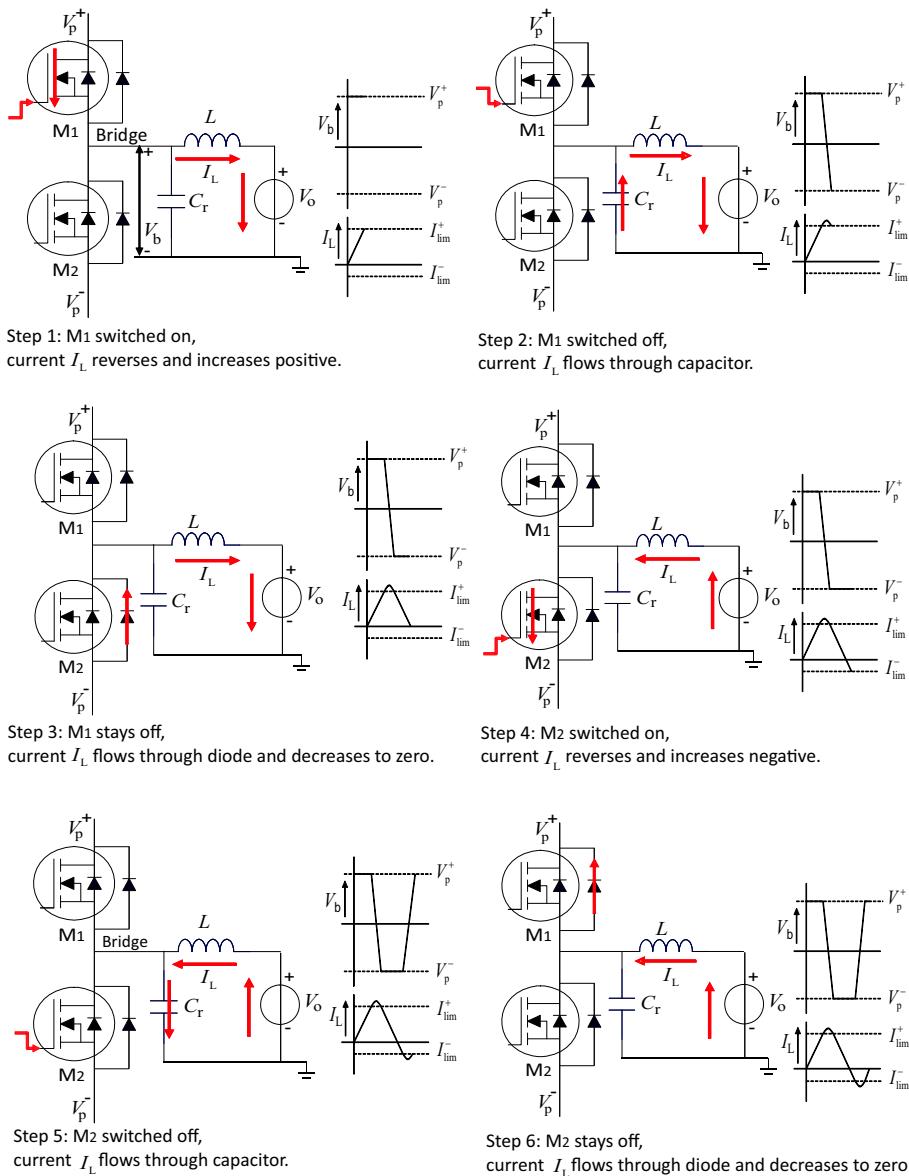


**Figure 6.78:** The controller of a resonant power output stage measures the current through the inductor. When this current has reached a certain limit, that is defined by the output current plus an offset, the MOSFETs are switched off. When the current is zero, the MOSFETs are switched on. This results in a controlled output current.  
(Courtesy of ASML)

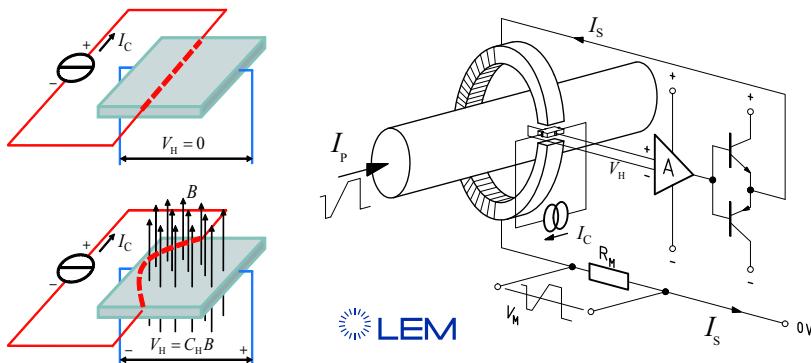
The switching behaviour with this controller is shown in Figure 6.79. In this figure some simplifications are applied. First the two capacitors from the bridge point to the power supplies are replaced by one capacitor to ground. This is allowed as power supplies have a voltage source characteristic with an equivalent impedance to ground of  $0 \Omega$ . The second simplification is the replacement of the capacitor of the output filter by a voltage source. This is allowed as the voltage over the capacitor will change only very little within one switching cycle of the high-frequent current, as was demonstrated with the normal switching output stage in Figure 6.72.

The main difference with the switching cycle of Figure 6.72 is the moment that the MOSFETs are switched off. This is determined uniquely by a positive current level for M1 and a negative current level for M2. After the MOSFETs are switched off, the voltage  $V_b$  of the bridge will not change extremely fast, because of the additional capacitor  $C_r$ . The slope is determined by the current level and the capacitor value. During this slope the energy in the inductor flows partly into the capacitor and the total behaviour is determined by the capacitor and the inductor only.

In fact it is this short moment in the natural resonating mode of the LC-combination that gave the name of “resonant-mode” to this configuration.



**Figure 6.79:** In a resonant switching power stage after switching one MOSFET off, the current is taken over by the capacitor and without any dissipation the bridge-voltage changes in a controlled way. The MOSFETS are switched on as soon as the current reverses.



**Figure 6.80:** A Hall sensor and a current sensing module that measures the current indirectly by measuring its related magnetic field with a Hall sensor. (courtesy of LEM)

The second difference is that the MOSFETs will only be switched on, when the current reverses in the branch, consisting of the MOSFET and his parallel diode. That switching moment effectively cancels all dynamic issues in either the diodes and the parasitic capacitors. In principle this alleviates the need for the additional parallel diodes but imperfections in timing will always remain so they are retained for additional robustness of the design.

One additional interesting observation can be done with this current-controlled power stage. Like with a normal switching output stage the duty-cycle of the switching needs to relate to the output voltage. This is true, because the output voltage of the LC-filter equals the average of the high-frequency signal before the filter and this average is determined by the duty-cycle. As a consequence of the current source behaviour of this output stage, the duty-cycle becomes a function of the load. In Figure 6.79 the duty-cycle is 50 % and with a short-circuited output of 0 V, this duty-cycle remains unchanged, even when the average current is not zero. The situation with a resistive load was shown in Figure 6.78. In that case both the duty-cycle and the frequency change.

The detailed description of all properties and signals in such an amplifier would go far beyond the goal of this book, unfortunately, but it is interesting to see how the current sensing can be done in such an amplifier.

### 6.3.4.2 Lossless current sensing

In very high power positioning systems, where these resonant power amplifiers are applied, the voltage levels are very high up to almost 1000 V, with current levels of several tens of Ampères. The simple inclusion of a series resistor to measure the current, would become a bit of a hazard. Furthermore, the three-phase configuration, that will be presented in the next section, requires the current to be measured at the high voltage output of the amplifier. For this reason a lossless current sensor is used that works with magnetism and active control.

The working principle of this current sensor is shown in Figure 6.80. It consists of a ferromagnetic ring with a small air gap, that is placed around the wire of which the current has to be measured. A Hall sensor measures the magnetic field in the air gap, that is generated by the current  $I_p$  in the wire.

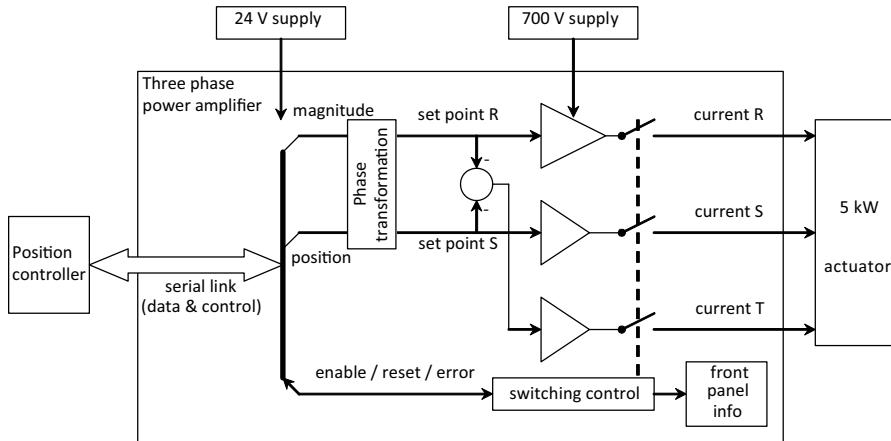
The Hall sensor is named after the American physicist Edwin Robert Hall (1855 – 1938), who discovered the *Hall effect* that determines the working of this sensor. A current  $I_c$  running through a conductive plate would normally run straight from one terminal to the other. The Hall effect is related to the influence of a magnetic field orthogonal to the surface of the plate as the Lorentz force on the current drives the electrons to one side of the plate. This results in an electromotive force that is measured as a voltage between two terminals, perpendicular to the current terminals.

The voltage of a Hall sensor is approximately proportional to the magnetic flux density  $B$ .

Because of the non-linear behaviour of the Hall sensor, a compensation circuit is applied, that consists of an auxiliary winding on the ring and an amplifier with a push-pull output stage. The current  $I_s$  through the auxiliary winding is controlled in such a way, that the magnetic field through the sensor remains zero. Because of the magnetic coupling of the auxiliary winding and the current conducting wire by the ferromagnetic ring,  $I_s$  is proportional to  $I_p$ .

### 6.3.5 Three-phase amplifiers

The need for ever higher accelerations and velocities of advanced mechatronic equipment has resulted in linear motors based on the Lorentz principle with three-phase electronic commutation, that have a large range of



**Figure 6.81:** Functional diagram of a three-phase power amplifier. The information from the position controller is transferred with a digital link to the phase transformation unit, that translates the information in a two phase (R,S) set point for the amplifiers. The third phase (T) is derived by subtracting the R and S from one. Switching current amplifiers deliver the current to the high power actuator.

(Courtesy of ASML)

motion. An example of such an actuator was presented in Section 5.2.5 of Chapter 5 on electronic commutation.

Generally, actuators in mechatronic positioning systems need to deliver a force that is only determined by the position controller and not by the position. It was demonstrated that a position independent force can be realised, when the three coil segments of a three-phase actuator are driven by an amplifier with three current outputs of which the magnitude of the current at a certain force set point changes sinusoidal as function of the position, in phase with the  $B\ell$  force factor of the related coil segment.

### 6.3.5.1 Concept of three-phase amplifier

The basic scheme of the second amplifier example of this chapter, that was designed for driving a 3 phase motor with 5 kW power, is shown in Figure 6.81. The digital controller provides a force setpoint via a serial data link to the amplifier, while status information is transferred back to the controller for diagnostic purposes. The force setpoint consists of the magnitude and the position of the actuator as this is necessary to guarantee that the currents are in phase with the actual  $B\ell$  force value of each actuator

coil segment.

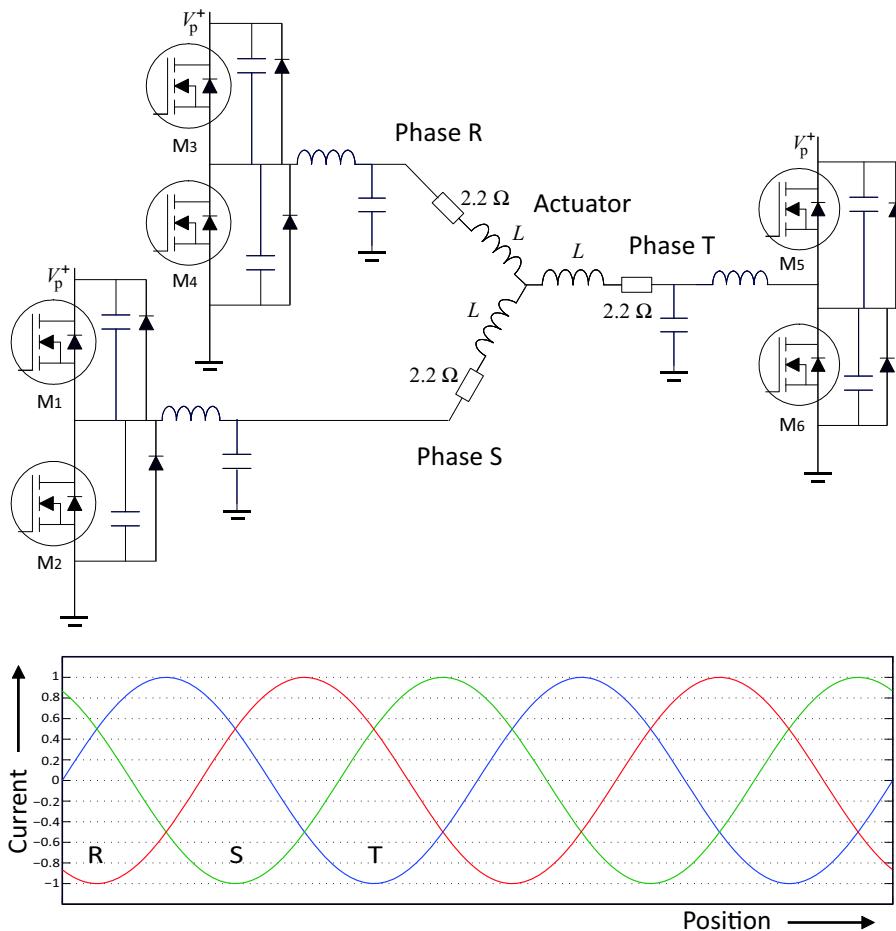
With this information a mathematical phase transformation is executed that creates two  $120^\circ$  shifted set point signals R and S with a sinusoidal relation to the position. Because  $R+S+T=0$ , the third setpoint signal can be determined by subtracting R and S from one. The three setpoint values are used as inputs for a current amplifier where, after digital to analogue conversion, each phase of the actuator is driven by a single-ended current-controlled power stage with a single power supply.

### 6.3.5.2 Three-phase switching power stages

Figure 6.82 shows how the three-phase actuator is connected to the three high-power resonant-mode switching output stages. In this configuration the three phases of the actuator coils are connected in a *star configuration*, with one central connecting point and the other terminals of the coil sections are connected to their corresponding phase terminal of the amplifier. Because of the application of only one power supply, this amplifier becomes a special version of the dual-ended amplifier from Section 6.3.3.8. The central point will have a voltage of an almost constant level, equal to approximately half the supply voltage. This voltage is constant as all currents at this node add to zero and the first of Kirchhoff's laws is fulfilled.

### Charge-pumping with three-phase amplifiers

One additional advantage of using three-phase amplifiers is, that the same rule of  $R+S+T=0$  means that at any moment in time a positive low-frequency current in one phase is compensated by the currents in the other phases. This means also that the charge-pumping from one phase is compensated by the charge-pumping of the other phases and as a result the net value of charge-pumping is zero.



**Figure 6.82:** A three-phase actuator only needs three single-ended output stages. Charge-pumping is avoided because all currents flow between the positive and negative power supplies, just like with the dual-ended configuration.

### 6.3.6 Some last remarks on electronics

The presented overview of power amplifiers is only so far complete, that it covers the most important effects of the main principles in combination with their application. For moderate power delivery and situations where power dissipation is not a big issue, linear amplifiers are preferred, because they avoid the related phase delay, caused by the output filter. For high power and situations where four-quadrant operation is necessary, a switched-mode amplifier is in fact the only solution.

Although not specifically mentioned anymore, a switched-mode amplifier has no problems with four-quadrant usage, as the power dissipation, due to the difference between the output voltage and the supply voltages, is avoided by the switching process and reactive components. Nevertheless the kinetic energy of a decelerating fast moving actuator needs to be dumped somewhere. When that energy is not converted into heat by dissipation in the amplifier, this has to be done in the power supply. The only way to achieve true high efficient four-quadrant power delivery to an actuator is by a combination of a four-quadrant power supply and a switched-mode power amplifier.

This methodology is one of the factors in the high efficiency of hybrid cars, where rotary actuators are driven by four-quadrant amplifiers and the batteries serve as four-quadrant power supplies to recuperate the kinetic energy from the braking actions. A modern hybrid car is indeed a real full mechatronic positioning system, although the precision is not extremely high.

# Chapter 7

## Optics in mechatronic systems

From the moment that mechatronics became a real precision engineering discipline, the application of optics was indispensable for achieving its ultimate performance. When controlling positioning accuracies in the order of a micrometre or less, reliable measuring tools are a prerequisite and only optical sensors can give the required combination of resolution, range and accuracy. But this is not the only relevant field of optics in respect of mechatronics. Modern imaging systems like telescopes, wafer lithography exposure systems and modern photographic equipment all need advanced mechatronic systems to control the imaging properties in both static and dynamic circumstances. This means that optics represent both a supporting technology and an application area for mechatronics.

Within its limited boundary conditions, this chapter will mainly focus on the understanding of several important terms and properties, related with the use of optics in mechatronic systems. It should help to enable the mechatronic designer to integrate optical systems in their designs in close cooperation with optical specialists. Where deemed necessary, based on our experience, the material is treated more in depth.

The chapter starts in Section 7.1 with the general properties of light and light sources. It will introduce the different properties of light sources like wavelength, coherence and radiance.

Section 7.2 introduces the physical model for *reflection* and *refraction* according to “Fermat’s principle”.

Section 7.3 deals with *geometric optics*. This is a practical method used to calculate the imaging properties of an optical system, where the light is approximated by ideal rays (ray tracing) and only the properties of materials on reflection and refraction are taken under consideration. This method is useful for a mechatronic engineer when designing simple optical systems that might for instance be used in laboratory test setups, made from “off the shelf” components.

For precision mechatronics the *physical optics* from Section 7.4 are the most important. Physical optics deals with the effects that are based on the wave character of light, including *interference* and *diffraction*. These effects limit the ultimate imaging properties of an optical system and are also used for position measurement with interferometers and encoders, as presented in the next chapter.

Section 7.5 gives a short introduction on adaptive optics, where active control is used to correct imaging errors, enabling near-to-perfect imaging properties of for instance terrestrial telescopes.

## 7.1 Properties of light and light sources

In physics the properties of light are described in two ways, as electromagnetic waves and as particles. The still mostly applied theory states that light acts like an electromagnetic wave, as was presented in Chapter 2. This theory explains the occurrence of interference and most of the observations that can be done with light.

The more recent theory on quantum optics is based on observations with extremely low energy levels of the light. The invention of the photomultiplier, working with a very high voltage, made it possible to detect light at these very low energy levels, by creating an avalanche of electrons as soon as one electron is excited by the light. It appeared that below a certain energy level, light starts to behave like separate particles called *photons*, each with a fixed energy that appeared to be equal to:

$$E = h f_p = \frac{hc}{\lambda} \quad [\text{J}] \quad (7.1)$$

with:

$h$  = Planck's constant ( $\approx 6.6 \cdot 10^{-34}$ ) [Js]

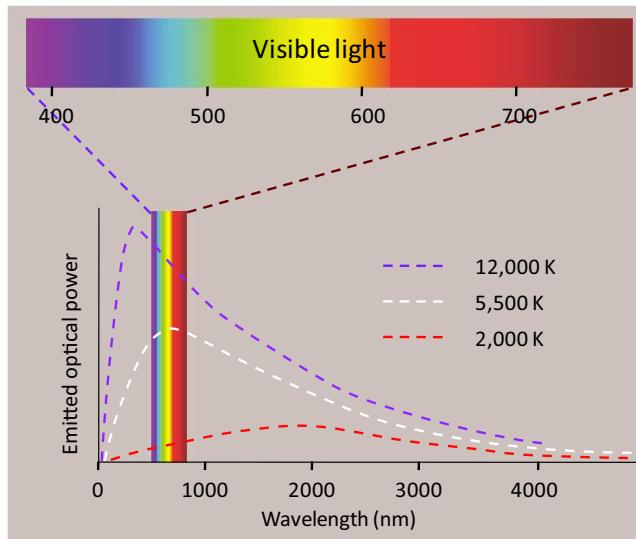
$c$  = The speed of light ( $\approx 3 \cdot 10^8$ ) [m/s]

$f_p$  = The frequency of the photon [Hz]

$\lambda$  = The wavelength of the photon [m]

When filling in some real numbers, the extremely low number of roughly  $E = 3.5 \cdot 10^{-19}$  J is obtained for 1 photon of visible light ( $\lambda \approx 600$  nm). A low level of one  $\mu\text{W}$  ( $\mu\text{J/s}$ ) of visible light power would hence be equal to about  $3 \cdot 10^{12}$  photons per second. This still immensely high number implies that it is not always necessary to take this quantum character into account. Nevertheless, with very sensitive measurements, the noise level is limited by the photon character and it partly explains why a larger image sensor in a photographic camera shows a better noise performance.

At several occasions it will appear to be practical to use the photons and electrons for the explanation of the physical phenomena that are related with light.



**Figure 7.1:** Thermal radiation spectra of a black body at different temperatures.

The visible light spectrum shows to cover only a small area of the wide spectrum of this kind of light source, explaining the low efficiency of incandescent lamps.

### 7.1.1 Light generation by thermal radiation

Many sources of light exist, ranging from thermal radiation by heated bodies to secondary emission of photons by electrons that change momentum or change state within an atom. Thermal radiation is often related to the visible light spectrum and is caused by the increasingly fast vibrations of charged particles inside atoms at increasing temperature. These vibrations of charge create electromagnetic waves, following the Maxwell equations. Due to this thermal nature of origin, the induced light consists of a wide range of wavelength values with a peak at a wavelength, which is determined by the absolute temperature in °K according to “Wien’s displacement law”, named after the German Physicist Wilhelm Carl Werner Otto Fritz Franz Wien (1864 – 1928). He stated that the shape of the wavelength distribution of a *black body*, defined as having a surface with a frequency independent emissivity, is independent of the temperature, while the maximum emission occurs at a wavelength equal to:

$$\lambda = \frac{b}{T} \quad [\text{m}] \quad (7.2)$$

With  $b$  being the Wien's displacement constant of  $\approx 2.9 \cdot 10^{-3}$  [mK]. With this relation a pure black body starts radiating visible red light at a temperature of about 2000 K. This corresponds with a peak radiation in the wavelength area of about 1500 nm, so most of the light is still in the invisible infrared region. At higher temperatures the peak emission will move to shorter wavelengths, while a black body temperature of 5500K has its peak emission around 530 nm, which equals the mid of the visible spectrum, like the daylight as emitted by the sun. A higher temperature will shift the peak to even shorter wavelengths into the invisible ultraviolet region. Figure 7.1 shows the spectra of these three examples.

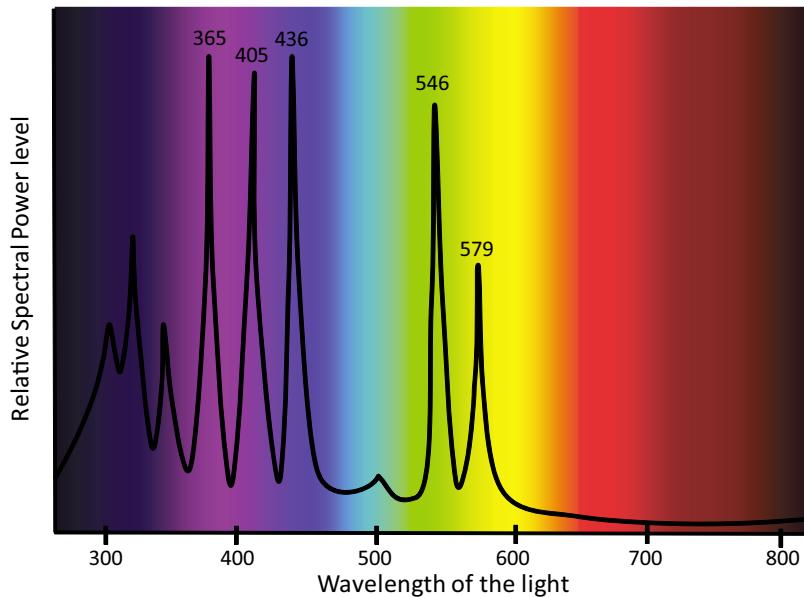
## 7.1.2 Photons by electron energy state variation

The aforementioned spectra from a thermal radiator show an inherently very wide range of frequencies. Often this is not preferable for technical applications, because many properties of optical elements are wavelength dependent. This leads for instance to the separation of sunlight into different colours in the raindrops of a rainbow that is caused by the wavelength dependent refractive index of water.

For that reason monochromatic light sources are often preferred. Fortunately several alternative light sources exist, mostly based on the emittance of photons by electrons as these change their energy state in a well defined way. Electrons in an atom can possess different energy levels, also called excitation states. Normally these electrons are in the "ground state", the lowest energy level. When excited by some external energy source, they move to different higher energy excitation states, depending on the energy that caused the excitation. These higher excitation states are generally not stable and the electron will ultimately fall back to a lower and more stable state or to the ground state, by emitting a photon of which the frequency is determined by the difference in energy levels of the electron states.

Electrons can be excited by temperature like in plasma gas discharge lamps, by fast moving other electrons like in fluorescent tubes or by means of other photons like in a laser.

The gas in a plasma gas discharge lamp, like used in the modern Xenon headlights of a car, is ionised by a very high voltage. As a result of this ionisation, part of the electrons will be able to move freely, thereby making the gas conductive and a current will flow. The corresponding electric power will heat the gas and it becomes a plasma with many free electrons that form a relatively low resistance path for a high current to run at a lower voltage of  $\approx 100$  V, than the value of  $\approx 1000$  V that was necessary to start



**Figure 7.2:** The light power spectrum of a mercury arc gas discharge lamp shows peaks corresponding with different energy states of the electrons around the mercury atoms in the plasma.

the process. These plasma gas discharge lamps radiate light at very specific wavelengths, determined by the gas atoms inside the plasma. An industrial mercury gas discharge lamp radiates with a spectrum as shown in Figure 7.2 with different narrow peaks of which the peak at 365 nm (I-line), at 405 nm (H-line) and at 436 nm (G-line) were used in the illumination of the early versions of wafer steppers.

The well-known fluorescent tube is also a Mercury type gas discharge lamp that works a bit differently. It does not work with a real plasma but the electrons are emitted by heated electrodes at the start and move to the other electrode through vacuum that is slightly doted with mercury atoms of which the outer electrons will be excited by the moving electrons. The 230 V AC voltage of the mains is not sufficiently high to start the lamp. For this reason a coil with a large self-inductance is used to both create the high starting voltage by interrupting the current through the coil by a “starter” switch and to heat the two electrodes by this starting current, in order to emit the first electrons that have to start the process. The peaks at 546 and 579 nm in the visible light region of Mercury are not sufficient to give an acceptable light spectrum for domestic use and most of the light is emitted in the invisible ultraviolet frequency range. For this reason a fluorescent

tube is coated at the inside with special fluorescent powder that converts the UV light into a wider spectrum in the visible light range, increasing both the efficiency and improving the colour. Although widely used in general lighting in buildings, these fluorescent tubes are of limited importance as light sources for precision mechatronic purposes due to their low radiance, a term that will be defined later and is related to the large surface that radiates the light with these lamps.

### 7.1.2.1 Light emitting diodes

Much more interesting for the mechatronic engineering field is the laser, which is the subject of the next section, but also the Light Emitting Diode (LED) is shown to be of increasing importance. LEDs are already used for many years as indicator in electronic equipment and only in the last few years they became known for illumination purposes. The principle of working of an LED is also based on electrons that change their excitation state, but now not within a gas but in a solid material. For that reason LEDs are also called *solid-state lamps*.

An LED is a special kind of semiconductor diode, of which the working principle was presented in Chapter 6. P and N material are combined causing either a barrier or a conductive path for the electric current, depending on the direction of the voltage.

In case of conduction, the electrons recombine with the holes at the transition zone from n to p material. This recombination implies the filling of an empty place around an atom, which corresponds to a change of the energy level. This would result in emission of a photon, were it not that Si or Ge are *indirect band-gap* materials, showing no radiation at recombination. This means that a normal diode will not emit light nor require the power that corresponds with that light emission. While this is not a problem for “normal” semiconductor diodes it does not create light. Fortunately other materials, composed from materials of adjacent valence bands (III and V) of the periodic system like Ga,As,In,P and N, have a *direct band-gap* and they do result in diodes that emit photons with a wavelength depending on the material, at the cost of additional electric energy. For instance GaAs gives infra red light, AlGaAs gives red light and GaAsP gives yellow light.

It has long been quite difficult to create shorter wavelengths, because that requires a wider band gap and not all material combinations are easily to manufacture. Eventually GaN appeared to be the right combination to create blue light.

The single colour of an LED and its rather low power of initially only a few

mW has long prevented the application in high power illumination systems, like domestic lighting. The breakthrough in creating blue LEDs made it possible to either combine red green and blue LEDs to create white light or use a phosphorescent material, to convert the short wavelength of the blue light to longer wavelengths. Much research has been spent on improving the efficiency and power output of both the LEDs themselves and on the phosphorescence, resulting in the recently introduced LED lamps for illumination purposes.

The main advantage of a LED as a light source in engineering applications is its almost monochromatic spectrum and its relatively small source dimensions. This will prove to be important for high resolution imaging as will be shown later when presenting the term *radiance*. LEDs are used for instance in the incremental optical measurement systems (encoders) that will be presented in Chapter 8.

### 7.1.2.2 Laser as an ideal light source

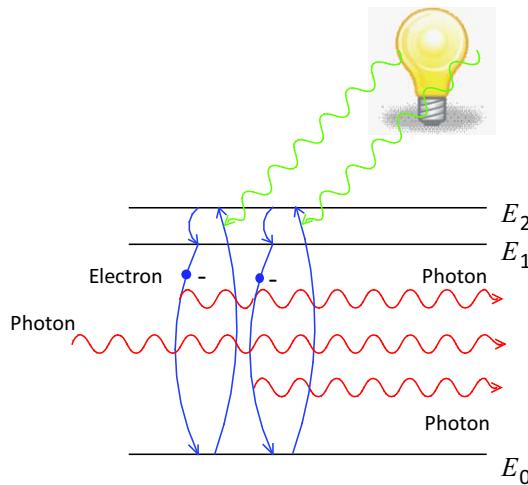
The ultimate light source is the *laser*. Laser is an acronym for Light Amplification by Stimulated Emission of Radiation and is based on the phenomenon that an electron at its excited state can be stimulated to drop to its lower state by a photon. This photon needs to have the same frequency as the frequency of the photon that the electron would emit by its change of energy state. In other words, one photon will create another identical photon by “triggering” an excited electron to drop to a lower energy state.

Like shown in Figure 7.3 one photon could stimulate two or more electrons resulting in three or more photons **with the same frequency and phase!**. This phase relationship is called *coherence*.

This physical process would of course come to an end, when all electrons have reached their lower state  $E_0$ . The process of emitting photons, when lowering the energy level of an electron, is fortunately reversible. This means that another light source, not necessarily coherent, like a normal gas discharge lamp, can be applied to excite the electrons to an energy level  $E_2$  **above** the excitation state  $E_1$  needed for the stimulated emission.

This higher energy level  $E_2$  is even less stable than  $E_1$ , so the electrons will drop to  $E_1$ , while sending a photon of a frequency that is not used in the lasing and can be neglected in the considerations. The additional external light source for exciting the electrons is called an “Optical Pump”.

It can be concluded that the complete system amplifies the incoming photons with a gain that depends on the frequency of the photons. This frequency is not infinitely sharply fixed, but has some tolerance due to uncertainty



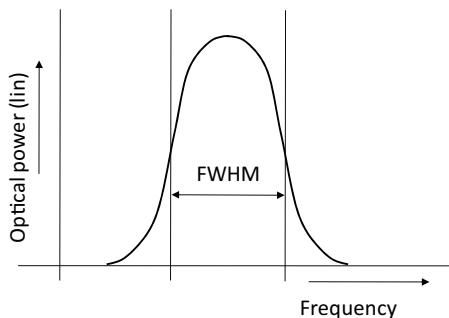
**Figure 7.3:** Light Amplification by Stimulated Emission of Radiation (LASER). An external source of energy excites the electrons to a higher unstable state  $E_2$ , where they first drop to a more stable state  $E_1$  until a photon with a suitable wavelength passes and triggers the electron to fall back to its original ground state  $E_0$ , under emission of a photon with the same wavelength as that of the trigger photon.

considerations on the energy levels of the electrons of which the deep treatise exceeds the scope of this book.

To accommodate this effect, the term “gain bandwidth” is used, which is quite different from the gain-bandwidth product that was defined for an operational amplifier. Optics designers often use the term *Full-width at half-maximum* (FWHM), the width of the frequency band that covers half the optical output power, as shown in Figure 7.4. This corresponds with the well-known -3 dB point on a logarithmic scale of electronic engineers.

The gain bandwidth tells something about the capability of the laser to amplify photons of a specific frequency within the bandwidth range and it automatically determines the bandwidth of the emitted light spectrum from a laser. Other words in literature for this quality are “line width” of a single frequency laser and “optical bandwidth”.

The bandwidth of a laser is one of its determining factors in the suitability as a light source for technical applications. For imaging purposes a high bandwidth can result in chromatic aberrations due to the frequency dependent properties of many lens materials. On the other hand a very small bandwidth can give problems due to the occurrence of scattered *speckle*, caused by interference. This phenomenon is known from the light of a laser

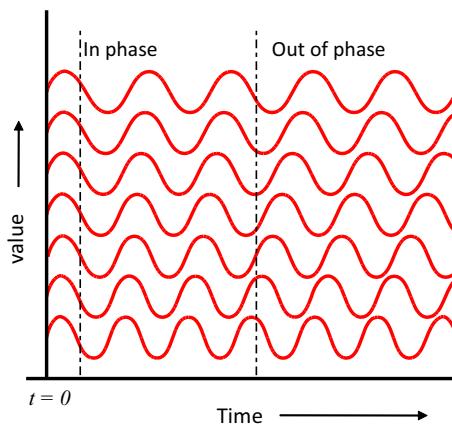


**Figure 7.4:** Definition of the full width at half maximum power gain bandwidth of a laser light source. The smaller the bandwidth is, the better is the temporal coherence of the laser.

pointer which shows tiny spots over the illuminated area that change as a function of displacement and angle of the incident light. These spots could impair the image of an object that is illuminated by a laser.

An other important property of a laser, partly related to this bandwidth and the occurrence of speckle, is the coherence. Two kinds of coherence exist namely *spatial coherence* and *temporal coherence*. The spatial coherence relates to the correlation between the electromagnetic field inside the laser beam at a certain moment in time **as function of the location in the beam**. It is a measure for the perfection of the shape of the emitted wavefront, a term that is explained later. The spatial coherence determines the divergence of the laser beam and hence the capability to keep a fixed diameter over a long range.

The temporal coherence of a laser relates to the correlation between the electromagnetic field on a certain location inside the laser beam **as function of time**. It is related to the bandwidth of the light from a laser and decreases as a function of the distance from the laser source. The correlation between the temporal coherence and the bandwidth is illustrated in Figure 7.5 where 7 graphs are shown of frequencies with a little difference between each frequency. In case all frequencies are in phase at the light source, which means that all values of the field are equally zero at  $t = 0$ , then the coherence disappears at some distance from the light source, at a time  $t = t_c$  that relates to the distance with the propagation speed of light. This time depends on the bandwidth of the source. With a wide bandwidth, the time is very short and with a single frequency source, this time is very long. The time that passes, until coherence is lost, is called the *coherence time*. Because light propagates at a constant speed in vacuum, this coherence time

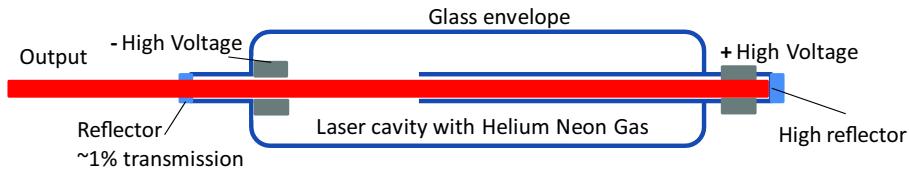


**Figure 7.5:** Electromagnetic field waves as function of time for light of different frequencies, starting all in phase. It shows the loss of phase relationship after a certain propagation time that corresponds to the distance from the light source.

directly relates to the more well-known and for technical applications more useful term *coherence length*, being the coherence time times the velocity of light.

In laser interferometry measurement systems as presented in the next chapter, the required coherence length is often several metres. As an example to get a feel for the required bandwidth of the light, a coherence length of ten metres implies that the lowest frequency in the bandwidth will need to be less than one period different from the highest frequency in the bandwidth over this length of ten metres. One period of for instance 500 nm relates to 10 m as  $5 \cdot 10^{-8}$ , so the bandwidth can not be more than  $5 \cdot 10^{-8}$  times the frequency.

This kind of precision is not achieved by the electron-photon physics in the laser only. Lasers, as used for high precision measurement systems, often consist of a combination of a laser tube with an optical resonance chamber, according to the Fabry-Perot principle as will be presented in Section 7.4.2.1. This resonator consists of two parallel mirrors at a very stable distance, such that a standing wave occurs between the mirrors. An example of a laser utilising this principle is the Helium Neon (HeNe) laser as shown in Figure 7.6 which radiates red light at 632.8 nm. This type of laser is already for many years used for distance measurement purposes, because the wavelength can be additionally stabilised by means of Iodine gas that shows a very narrow natural frequency and enables active control of the



**Figure 7.6:** Cross section of a Helium Neon Laser. A high voltage creates a plasma where the electrons of the gas are continuously excited to their higher energy state and emit photons of different wavelengths, while falling back from that state. Lasing takes place induced by photons of only one wavelength, leading to a maximum emission at 632.8 nm. The two reflectors create an optical resonator that further enhances the frequency stability of the laser.

distance between the mirrors. It enables the use of this laser as a reference for the standard metre. Next to this precision HeNe-laser, a large amount of different laser types are introduced of which the semiconductor diode-laser has been instrumental to the optical registration of signals on a disk like the red laser diode for the CD-player and the blue lase diode for the Blu-Ray player. These diode-lasers work according to the same principle as an LED but with additional internal lasing and small mirror facets on the exit points of the semiconductor material to create a Fabry-Perot resonance chamber.

### 7.1.3 Useful power from a light source

Besides the frequency and phase of a light source, also the radiated energy, the size of the source and the direction of radiation are of importance. These all influence the amount of light that can be captured by an optical system and transmitted to the area of interest.

The *illumination power*  $P_i$  [W] of the radiated light, sometimes also called the *radiant flux* [ $\Phi$ ], is not sufficient to define the usefulness of a light source. First of all it contains all frequencies. But even when a light source is monochromatic, the pointing direction, spread and the radiating surface are undefined. Many terms are introduced to narrow down these undefined parameters of a light source in relation to the optical system, of which the following are explained here:

- The radiant emittance and irradiance.
- The radiance.
- The optical throughput or etendue.

### 7.1.3.1 Radiant emittance and irradiance

The *radiant emittance*  $M$  from a light source equals the illumination power, radiated by the source, divided by its radiating surface [ $\text{W/m}^2$ ].

The radiant emittance closely relates to the term irradiance that was defined in Chapter 2. The irradiance  $I_r$  equals the power density of the received, incident light on a defined surface [ $\text{W/m}^2$ ]. Often the term intensity is also used for this term but the definition is not very clear as in optics and radiology the term intensity is also used for the energy flow per unit of solid angle. For this reason the term irradiance is preferred in optics.

The irradiance determines all observations in optics as it is the only property of the light that can be measured with a sensor or the human eye. A photo detector with a given surface converts the light power, being the irradiance times the surface, into an electric current.

With a given value for the radiant emittance of a source, the irradiance, received at a surface in its neighbourhood, decreases proportional to the squared distance from the source. Contrary to the terms from the next part, the irradiance can be increased again by concentrating light on a spot with an optical system.

### 7.1.3.2 Radiance

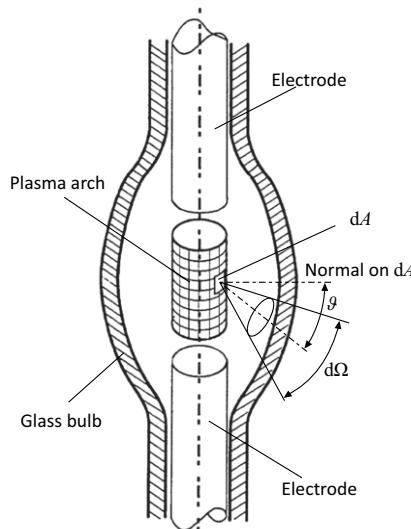
The size of a light source relates to its spatial coherence and for that reason a *point source*, with an infinitely small radiating surface, pointing into the right direction, appears to be the most ideal in respect to the performance of an optical system. In that respect, a perfect parallel laser beam fulfils that demand, as it can be modelled as a point source located at infinity, pointing in only one direction.

The term *radiance* is used to describe the extent to which a light source behaves like such an ideal point source. The radiance is proportional to the radiant emittance and for that reason it is also often called the *brightness* of a light source. As will be shown, the radiance also includes the radiated direction of the light. With a real point source, the radiance would by definition be infinite.

In physical terms the radiance  $L$ , see Figure 7.7, is defined within the next expression for the illumination power, radiated from an infinitesimal part of a light source in a certain direction:

$$dP_i = L(x, y, \alpha, \beta) d\Omega dA \cos \vartheta \quad [\text{W}] \quad (7.3)$$

With:



**Figure 7.7:** Definition of radiance of a light source. As example a plasma gas discharge lamp is taken, where the plasma arch is approximated as a light emitting cylinder. (Courtesy of ASML)

- $A$ : The radiating surface of the light source. [m]
- $d\Omega$ : The solid cone angle that defines the radiating direction of an infinitesimal radiating surface element  $dA$ . [sr]
- $dP_i$ : The illumination power, radiated from the infinitesimal surface element into its related solid cone. [W]
- $L(x, y, \alpha, \beta)$  The radiance of the source at the infinitesimal surface element located at coordinate  $x, y$  in the direction  $\alpha, \beta$  into its solid cone. [ $\text{W}/\text{m}^2\text{sr}$ ]
- $\vartheta$ : The angle between the surface normal and the direction of the radiating solid cone. [rad]

When the radiance is constant over the total surface  $A$  and the solid cone angle is pointing in the direction of the normal on the surface, the expression of the total illumination power of a light source into a certain direction, defined by the solid cone angle  $\Omega$ , can be simplified into:

$$P_{i,t} = LA\Omega \quad [\text{W}] \quad (7.4)$$

As previously mentioned, a small surface of the source appears to be better for the performance of an optical system. This is due to the physical law of

**conservation of radiance** stating that **the radiance**, as received by an optical system, divided by the refractive index squared, **can never increase**. At best the radiance remains the same over the optical path of a system<sup>1</sup>. The importance of this conservation law of radiance is illustrated with an example from photography. The energy of the light that is needed to activate one pixel at a digital camera appears to determine the minimum radiance level of the light at the source when a certain maximum exposure time is allowed. To understand this, it is first necessary to realise that an image in an optical system acts as a source of light for the next element, in this case the photo sensitive pixel in the sensor. That means that the total power of the image spot at that pixel can be calculated with Equation (7.4). The dimension  $A$  is defined by the imaged object and the solid angle  $\Omega$  is defined by the aperture of the photographic lens. The illumination power is determined by the necessary energy to activate the pixel, divided by the exposure time. These terms together determine the required radiance of the source:

$$L \geq \frac{P_{i,t}}{A\Omega} \quad (7.5)$$

Another example to underline the effect of this conservation law is observed, when imaging a large object like the sun by demagnifying optics like a burning glass. The image of the sun becomes a very bright spot with an increased irradiance. On the other hand however, the surface of the image is much smaller than the surface of the object. Corresponding with that demagnification, the solid angle of the light at the image is proportionally larger than the solid angle of the incident light rays, as these are almost parallel. As a result the radiance is not changed. A third example is based on the idea that the radiance of a light source might be increased by shielding a part of its surface. That method can be useful in case one needs a smaller radiating surface, but the radiated illumination power will decrease proportionally and again no gain in radiance is achieved. It can even be proven that concentrating the light from all surface elements of a light source onto a small hole, will not increase the radiance of the resulting spot, even though the irradiance is high.

---

<sup>1</sup>Normally the refractive index is the same at the source and the image. Only with immersion optics in lithographic exposure systems, the radiance can be increased at the image because of the increased refractive index.

### 7.1.3.3 Etendue

From the previous it can be concluded that the radiance is maximum when the light is radiated from a very small surface at a very small angle. Unfortunately only an ideal laser is capable of creating such a perfection and in reality the optical system can only collect a part of the light.

For this reason another quality term is introduced, the *optical throughput*, also often named with the French word *etendue*. The etendue is a measure for an optical system of its ability to collect as much as possible light from a source and retain as much as possible of the original radiance.

The full calculation of the etendue requires an integration of the light sent by all infinitesimal surface elements of the source and received by all infinitesimal surface elements of the entry-pupil<sup>2</sup> of the optical system. In the practical case of a light source with a small surface, relative to the entry-pupil of the optical system, where also both surfaces of the radiating source and the entry-pupil are essentially parallel ( $\theta = 0$ ), the etendue  $G$  can be approximated as:

$$G = \Omega A \quad [\text{m}^2\text{sr}] \quad (7.6)$$

With Equation (7.4) this results in the following approximating expression for the illumination power, captured from the source by an optical system:

$$P_{i,c} = LG \quad [\text{W}] \quad (7.7)$$

With a given amount of necessary illumination power and a given radiance of the source, this means that the etendue should be maximised. Corresponding to the law of conservation of radiance, this means that the etendue should be maximum directly at the entry-pupil, as it can not be **increased** by optical means **after the entry-pupil**. This effect is illustrated with the last example from above, about shielding a part of the surface of the source. This would decrease the etendue, resulting in a decrease of the illumination power.

---

<sup>2</sup>As will be explained later, the entry-pupil is the cross section of an optical system that defines which incident light rays are captured.

## 7.2 Reflection and refraction

To introduce the subject of this section it is necessary to remain aware once more that physical theory is based on observations that have to fit in a valid model. One of the first and most striking observations that people have made on light is that it travels along straight lines as long as the material, where it works in, has constant properties.

It was the French lawyer! Pierre de Fermat ( $\approx 1601 - 1665$ ), who postulated that light follows the path that requires the least time to reach a certain place. This **Fermat's principle of least time** is still used today and it will be used in this section to explain phenomena like reflection and refraction. Before entering in these phenomena it must be mentioned that this again is a strange theory, as one could wonder how light as a series of photons can know which path to follow to be first in time. Well like Richard Feynman states in his famous “Feynman lectures on physics”:

But what does it do, how does it find out? Does it smell the nearby paths and check them against each other? The answer is yes, it does, in a way.

To understand this a bit better it is necessary to revert to the theory of light, where a photon can also be represented as a continuous wave and this wave behaviour gives the photons their “information” about the path and the obstructions in the path. This is sufficient for the scope of this book to be able to use it in the next part to describe wavefronts.

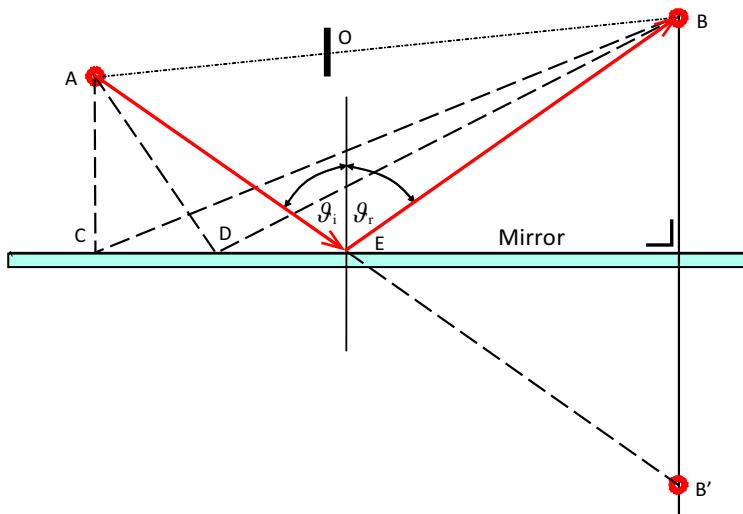
Like stated in Chapter 2, reading Richard Feynman’s work is highly recommended for those readers, who would like to find out more about this duality of light and the related theory that is based on a probability analysis on the path of photons that in principle can take any path but most probably will only follow the path of least time.

When presenting phase gratings, this theory will be used to explain mirrors that reflect in other directions than as presented in the next section.

### 7.2.1 Reflection and refraction according to the least time

In Figure 7.8 a light source **A** is shown that emits photons.

When determining the path of the photons to place **B**, according to Fermat they should follow the path of least time. As photons only move in straight



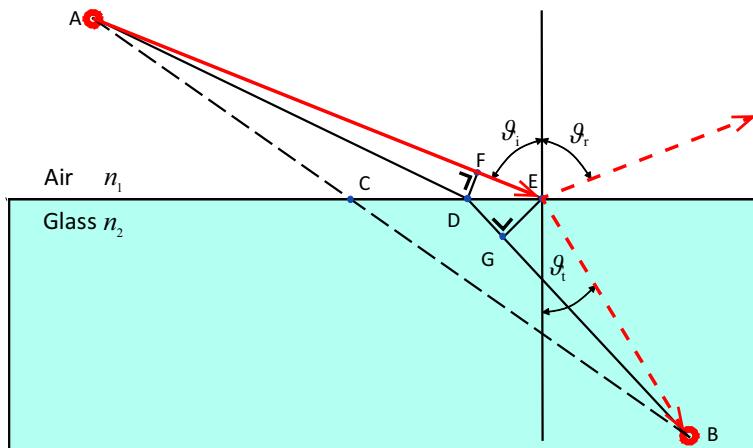
**Figure 7.8:** The path of the photons, reflecting from the surface of a mirror derived by the principle of least time, results in equal angles of incidence  $\vartheta_i$  and reflection  $\vartheta_r$ .

lines as long as they remain in the same medium, the shortest time corresponds with the shortest distance. This shortest distance is a straight line that can be drawn directly from **A** to **B**, but that path is obstructed by screen **O** and the photons can only arrive at **B** via the mirror.

Evidently the path via **C** is not taking the least time and also via **D** it is not optimal yet. To determine the exact right place to reflect, a little trick is used and a mirrored place **B'** is drawn. When drawing a straight line between **A** and **B'** it is known that this is the shortest path so, when the photon is reflected at the intersection point **E**, this is the shortest path to **B**. With a little bit of trigonometry it becomes clear that the angle of incidence  $\vartheta_i$  equals the angle of reflection  $\vartheta_r$ , which corresponds with what was taught at high school.

This relation is not too difficult to derive for simple reflection, but the real value of Fermat's model is in predicting the path, when the photons go from one medium to another. For this derivation Figure 7.9 is drawn, where photons from the same source **A** located in a low density medium (air) now have to reach position **B** located in a high density medium like glass.

First of all it was stated by Fermat that a more dense medium will slow down the light. The speed of light ( $c_m$ ) in any medium compares with the



**Figure 7.9:** The path of photons refracting at the interface between two different media, derived by the principle of least time, results in unequal angles of incidence  $\theta_i$  and refraction  $\theta_t$ . The shortest path is not the fastest! Note that also some light is reflected.

speed of light in vacuum ( $c$ ) with a factor equal to its refractive index ( $n$ ).

$$c_m = \frac{c}{n} \quad (7.8)$$

This means that it takes more time to propagate in a high density medium, than it would take in a low density medium for the same distance. To accommodate this difference in speed, the term *Optical path length* (OPL) is often used, where OPL equals  $n$  times the real geometric path length. In terms of time the method of Fermat is equal to stating that photons search for the shortest **optical** path, hence taking the refractive index into account<sup>3</sup>.

The shortest path from **A** would be a straight line intersecting the interface between the media at **C**. This geometric path does however not correspond with the shortest optical path with the least time, as the part from **C** to **B** would add considerable to the time.

One might say that the photons recognise that it is better to stay a bit longer in the air and find a more optimal path, the shortest optical path.

When moving the intersection point from **E** to the right, first the gain

<sup>3</sup>One might conclude that photons slow down in a high index medium, but photons always travel with the same constant speed. The observed delay is caused by interaction with the electrons in the medium, where the original photons are exchanged for new photons with a small delay.

in time from shortening the path from the intersection to **B** outweighs the loss of time from lengthening the path from **A** to the intersection. At a certain moment however this gain becomes zero and the total time will increase again. This can be simply understood, when considering an extreme intersection point, fully to the right in the figure.

Somewhere in between, the optimal point is found and as a first assumption point **E** is taken. An optimum means that the change of the optical path length as function of the location on the intersection becomes zero. This means that the time for following the optimal path, compared with the time for a nearby path, through for instance intersection point **D** need to be almost equal. By drawing two perpendicular lines, one from **D** to **F** and one from **E** to **G**, the difference in path length though air, when comparing intersection point **D** with **E**, is approximately the length between **F** and **E**  $\langle EF \rangle$ . The difference in path length through the glass is approximately the length between **D** and **G**  $\langle DG \rangle$ . This approximation is valid when distance between **D** and **E**  $\langle DE \rangle$  is very small. If **E** is the optimal point then:

$$\text{OPL}_{\text{air}} = n_1 \langle EF \rangle = \text{OPL}_{\text{glass}} = n_2 \langle DG \rangle \quad (7.9)$$

$\langle EF \rangle$  equals  $\langle DE \rangle$  times the sine of the angle between points **F**, **D** and **E**. From trigonometry it follows that this angle is equal to the angle of incidence  $\vartheta_i$ .

Likewise  $\langle DG \rangle$  equals  $\langle DE \rangle$  times the sine of the angle between points **E**, **D** and **G**, being equal to the angle of the transmitted and refracted light  $\vartheta_t$ . From this all follows:

$$n_1 \langle DE \rangle \sin \vartheta_i = n_2 \langle DE \rangle \sin \vartheta_t \quad (7.10)$$

After cancelling  $\langle DE \rangle$ , this equation gives the law of Snell, named after the Dutch mathematician and astronomer Willebrord Snel van Royen(1580 – 1626), also known with the Latin name “Snellius”. When light goes from a medium with refractive index  $n_1$  to a medium with refractive index  $n_2$  Snell's law becomes as follows:

$$n_1 \sin \vartheta_i = n_2 \sin \vartheta_t \quad (7.11)$$

When calculating the refracted angle from a known incident angle, Snell's law is written as:

$$\sin \vartheta_t = \frac{n_1}{n_2} \sin \vartheta_i \quad (7.12)$$

### 7.2.1.1 Partial reflection and refraction

The above reasoning from the least time also shows that reversing the direction of the photons, so going from glass to air will result in the same shortest path, followed in the reverse direction. In this reversed situation, the angle of incidence in the original law of Snell becomes the angle of refraction and vice versa, resulting for the refracted angle as function of the incident angle:

$$\sin \vartheta_t = \frac{n_2}{n_1} \sin \vartheta_i \quad (7.13)$$

Because  $\sin \vartheta_t$  can never be larger than one, above the *critical angle*  $\vartheta_i = \arcsin n_1/n_2$  no light will refract any more and total reflection occurs. This effect is known from light guiding in fibres and flat glass or acrylic plates, when illuminated from the side.

This effect does of course not occur only at or above the critical angle and in reality at any optical surface a combination of reflection and refraction occurs. For reflected light the irradiance ratio to the incident light is given in the following equation:

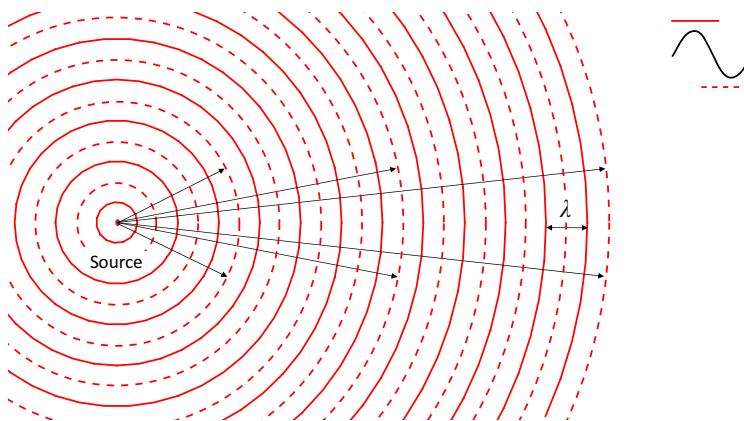
$$\frac{I_{r,r}}{I_{r,i}} = \frac{n_1 \cos \vartheta_i - n_2 \cos \vartheta_t}{n_1 \cos \vartheta_i + n_2 \cos \vartheta_t} \quad (7.14)$$

with  $n_1$  respectively  $n_2$  equalling the refractive index of the medium of incidence respectively the other medium.

Without the need to be able to derive this formula it shows for instance that with very small angles of incidence this ratio will be very small, which means that most of the light will be transmitted. At larger angles this ratio will become larger until most of the light is reflected.

The equation can be checked, because at a certain angle 100 % reflection can be achieved, when going from a higher refractive index to a lower value. This corresponds with the equation as  $\vartheta_t$  will become  $90^\circ$  at the critical angle  $\vartheta_i = \arcsin n_1/n_2$  and  $\cos \vartheta_t$  will then be zero resulting in one as the outcome of the equation.

In practical optics mostly a controlled amount of either reflected or refracted light is aimed for. This is achieved by coatings on the surface that consist of successive thin layers of materials with a different refractive index that form an optical resonator, like mentioned with the laser. Such a coating can be tuned to either reflect or transmit light with a certain wavelength or a range of wavelengths.



**Figure 7.10:** Wavefront model of the propagation of light. Concentric spheres represent photons that simultaneously originated at the source in the centre of the sphere. The distance between the spheres represents either the wavelength of the light or a proportional value thereof.

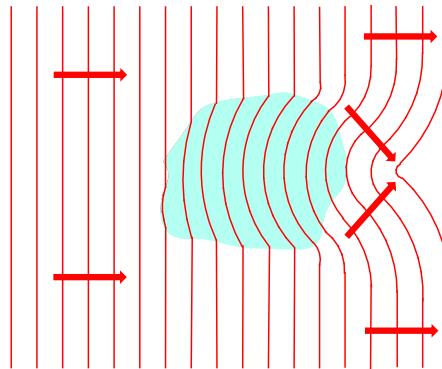
### 7.2.2 Concept of wavefront

Before applying these rules on reflection and refraction, first the concept of a wavefront is introduced. When observing a number of photons in vacuum, all originating from a light source at  $t = 0$ , these photons will move with the same constant speed of light in different directions away from the source. At any moment in time, they will all be at the same distance from the source and their position is defined by a sphere with its centre at the source.

The wave equivalent of this reasoning is as follows. At any moment in time, light from a monochromatic coherent light source in vacuum will be in phase on any concentric sphere around the source. For that reason these spheres are called a wavefront, which is a mixing of the motion character of the photons and the continuity character of the waves.

When working with this thinking concept of a wavefront, it is necessary to remain aware that it is a very artificial concept as in reality light never is completely monochromatic nor completely coherent over long ranges, like was explained with the laser. A coherence length of metres is large, so light from a star is never coherent. Nevertheless it is a very useful concept, because it visualises different optical effects.

It is used to illustrate phase relations and the local direction of the light is always orthogonal to the wavefront. A curved wavefront corresponds with diverging or converging light rays and a flat wavefront corresponds with parallel rays of light. This is illustrated in two dimensions in Figure 7.10,



**Figure 7.11:** Distortion of a flat wavefront caused by the reduced propagation speed of part of the wave, when passing through a medium with a higher refractive index.

where the spheres are approximated by circles. The waves originating from the point source are depicted, as if they have only one frequency with a wavelength equal to the distance of the solid or dashed lines, while the solid lines represent the positive maxima and the dashed lines the negative maxima of the sine wave.

Often the dashed lines are shown only when the phase relationship is important. When only the direction is investigated, they often are omitted.

### 7.2.2.1 A wavefront is not real

In reality the spheres would need to have a very small distance, equal to the wavelength of the light and they would need to move with the speed of light. This means by definition that a drawing of a wavefront only shows a snapshot at a distinct moment of a propagating travelling wave. It does by no means imply that a wavefront drawing represents a standing wave.

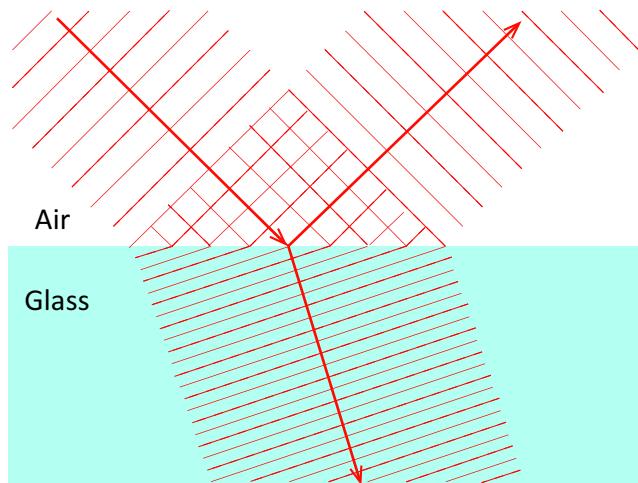
When examining interference at a very small scale, like in Section 7.4, the distance between the spheres is equal to the wavelength. At large scale optics however, when only the direction of the light matters, often drawings on wavefronts are used with much larger distances of the lines, where their distance is only proportional to the real wavelength.

From Figure 7.10 it is clearly observable that the wavefront becomes more and more flattened out with increasing distance from the source, which explains the almost parallel rays of light coming from the sun.

Several qualitative statements can be derived with the wavefront model

regarding the effects of changes in the refractive index of the medium where the waves pass through. In case these media are not homogeneous, the waves will partly slow down resulting in a distorted wavefront, like shown in Figure 7.11. In fact the medium in blue acts like a kind of lens and it illustrates the observed effects, when we look through non homogeneous gas or fluid, like the air above a heated surface.

In Figure 7.12 the previously described reflection and refraction is shown with flat wavefronts and a shorter wavelength (lower propagation speed) in the glass. From the fact that at the connection points both reflected and refracted waves should have a fixed phase relationship, the law of Snell can also be derived in a graphical way. On purpose, the term “fixed phase relationship” is used, instead of “the same phase”, as at reflection from a less dense medium to a more dense medium a  $180^\circ$  phase shift occurs, comparable with the rigidly connected chain of mass spring systems in Chapter 2. More details on this aspect can be found in specialised books, dealing with the Fresnel equations.



**Figure 7.12:** Reflection and refraction at the interface between two different media derived by using wavefronts. A higher refractive index causes a shorter wavelength that at the interface is in phase with the incoming wavefronts. The reflected wavefront shows  $180^\circ$  phase reversal, because of the reflection from a low density to a high density medium.

## 7.3 Geometric Optics

Geometric optics is an important area of optics used for predicting imaging properties of optical systems by approximating the light as an independent set of separate rays and by neglecting effects like interference and diffraction that relate to the wave character of light. This approximation is only valid, when the important details of the object to be imaged are essentially larger than the wavelength of the light, but even in more critical designs geometric optics can give very useful predictive results.

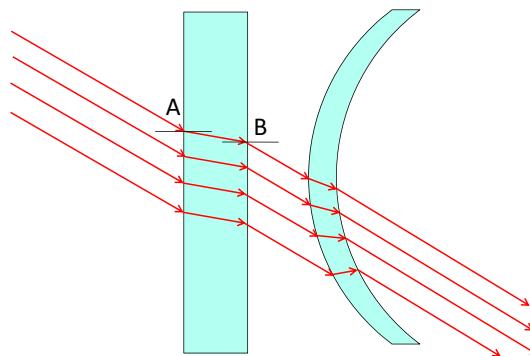
Optical imaging is based on the two main principles that were introduced in the previous section, reflection and refraction. In case of refraction, the transparent optical elements are called lenses and in case of reflection, the non-transparent optical elements are called mirrors. Reflective systems are also named *catoptric* systems and refractive systems *dioptric*, a term that relates to the term *dioptre* for the “strength” of a lens. The dioptre is the reciprocal of the focal length and mainly used within ophthalmology. A combination of lenses and mirrors in an imaging system is called a *cata-dioptric* system.

As most smaller optics from everyday life consist of refractive elements, this section will be limited to dioptric optics only. The treatise is in principle more complicated than with the catoptric counterpart, due to the refraction involved.

Two essential differences however need to be mentioned. First of all, refraction in general is wavelength dependent and as will be shown, this dependency introduces chromatic aberrations that do not occur with reflection. Secondly an imaging system with only refractive lens elements is less dependent on small rotations (tilting) of each element around the axes perpendicular to the optical axis, while any reflective surface is extremely sensitive for tilting. This property poses severe requirements on the mounting stability of these reflective elements.

### 7.3.1 Imaging with refractive lens elements

Parallel rays of light approaching a piece of glass with two parallel planar surfaces will be refracted and continue their path with the same parallelism and distance. This is the case, because the refraction at the point of insertion is compensated by the refraction at the exit point of the glass. This can be seen in Figure 7.13 that also shows what would happen with a piece of glass with two curved parallel surfaces with the same curvature. Even in that



**Figure 7.13:** Refraction of two pieces of glass with parallel surfaces, one flat and one curved, shows that only a shift of the lines will occur. The refraction at point **A** is reversed at point **B**

case the rays would propagate approximately parallel after the passage and only their distance will be changed.

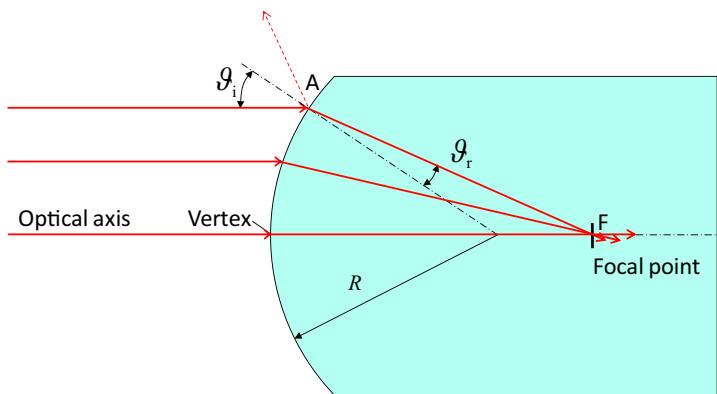
In order to achieve an optical image it is necessary to use a piece of glass with a varying thickness, slowing down the rays depending on their incidence angle and position.

The invention of a lens is already more than three thousand years old. By coincidence of nature it is far more easy to achieve a curved surface than a flat surface with optical quality. When polishing a piece of glass by a circular motion against a relatively soft counter surface with abrasive paste, automatically a curved surface will result. The abrasion takes place most at the outer part of the piece of glass, because the outer part makes a larger movement in this process and the force exerted near the edges is also larger, resulting in an essentially spherical surface.

As a first step into the modelling of the imaging properties of a lens, the refraction at a spherical surface with radius  $R$  is shown in Figure 7.14.

First the *optical axis* is defined as being the axis of rotational symmetry of the optical system. A ray parallel to the optical axis, entering the surface at point **A**, is refracted in the direction of the normal towards the centre of the surface, according to the law of Snell. A ray drawn at the optical axis would have an angle of incidence of  $90^\circ$  in the Vertex. As a consequence this ray would continue its path with the same direction and intersect the first ray at **F**.

It can be reasoned that all parallel rays at the same distance from the optical axis as the first ray at **A** will after refraction cross the optical axis at the



**Figure 7.14:** Refraction at a spherical surface results in converging rays that intersect at the focal point  $F$ .

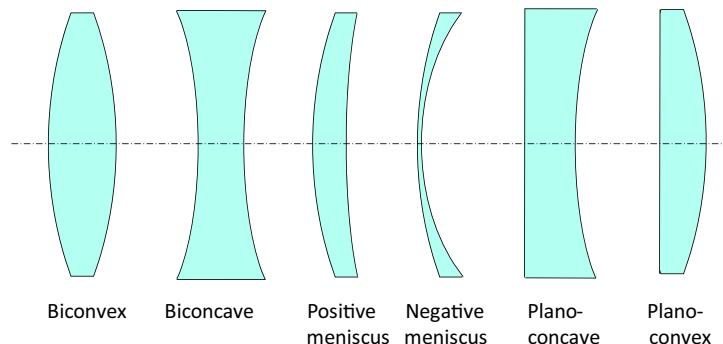
same point  $F$ , because of the rotational symmetry and the point  $F$  is called the focal point.

As a an approximation also rays entering closer to the optical axis will bend towards the same focal point  $F$  with only a small error causing the “spherical aberration” of a lens. This aberration is small, when only those rays are considered that are very close and almost parallel to the optical axis, the so called *paraxial rays*.

### 7.3.1.1 Sign conventions

It is important to introduce some sign conventions to achieve the right sign for the image, positive for the same side of the optical axis as the object and negative for the opposite side. The formal sign convention in imaging systems starts with the light source or object positioned **at the left side of the system** and is a bit counter intuitive, as it does not correspond with a standard coordinate system. The following is commonly used:

- $F_o$ , the distance of first focal point to lens is positive when located at the left side of the lens.
- $F_i$ , the distance of second focal point to the lens is positive when located at the right side of the lens.
- $S_o$ , the distance of object to lens is positive when located at the left side of the lens.



**Figure 7.15:** Different lenses exist with names that are determined by the curvature of their two surfaces.

- $S_i$ , the distance of image to the lens is positive when located at the right side of the lens.
- $y$ , the vertical axis is positive above the optical axis.
- $R$ , the radius of the surface of the lens is positive if the centre is located at the right side of the lens.

As a consequence of these conventions a biconvex lens has for instance one positive radius (left) and one negative (right).

### 7.3.1.2 Real lens elements

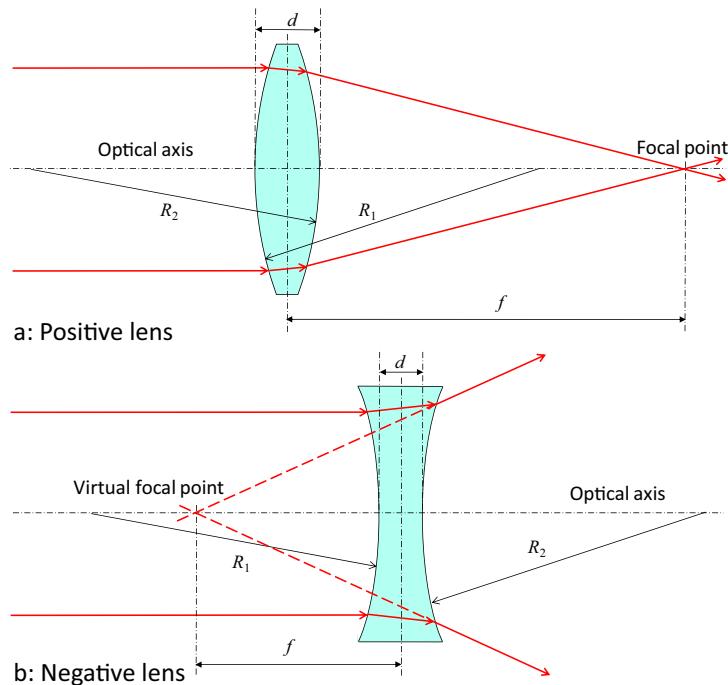
A real lens consists of two surfaces with different curvatures where the same law of Snell can be applied.

A wide variety of lenses exist, as shown in Figure 7.15. The different names are derived from the curvature shape. A rounded surface is either convex, flat (plano) or concave and all combinations can be applied.

A biconvex lens is called a positive lens as it will bend the rays towards the optical axis (converging). Consequently a biconcave lens is a negative (diverging) lens, and any other combination is positive or negative depending on whether the optical path through the lens is largest near the optical axis or at the outer part of the lens.

A combination of a convex and a concave surface is called a meniscus because of the similarity with the shape of the human meniscus and it is the standard configuration for eye-glasses.

The first example of imaging optics that will be shown in this section, is the positive biconvex lens from Figure 7.16. This lens refracts the rays in

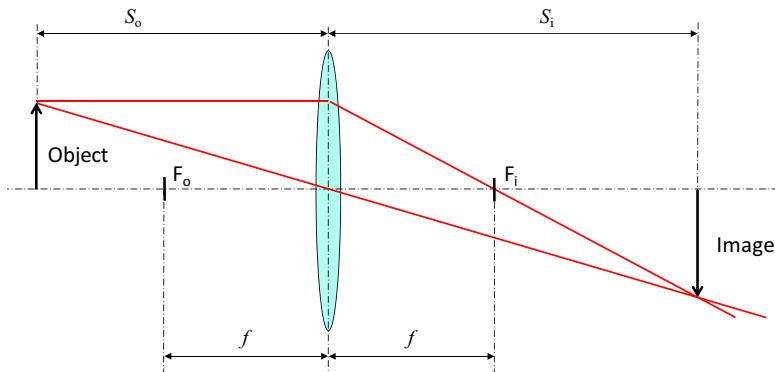


**Figure 7.16:** A positive lens, consisting of two convex surfaces that both converge the refracted rays to the optical axis, gives a shorter focal length, than with one convex surface only.

A negative lens, consisting of two concave surfaces that both diverge the refracted rays, creates a “virtual” focal point at the same side as the object.

a converging direction towards the optical axis. It can be seen that both surfaces refract the rays that come from the left in the same direction, resulting in a focal point closer to the lens than would be the case with one surface only. With a negative biconcave lens the opposite result is obtained. In this case, the rays are refracted away from the optical axis and seem to be originating from a point located at the left side of the lens. This non-existing point is called a *virtual* focal-point, because it is not the real origin of the rays that are observed at the right side of the lens.

The same method of “the least time” by Fermat can be used to derive the equation for the focal length of a lens but it is also possible to use the law of Snell directly.



**Figure 7.17:** A positive thin lens gives a real image, when the object is further away than the focal length  $f$ .

Without showing the total derivation, in both cases the following equation can be derived for the focal length of the lens:

$$\frac{1}{f} = (n - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} + \frac{(n - 1)d}{nR_1R_2} \right) \quad (7.15)$$

with:

$f$  = the focal length of the lens [m]

$n$  = the refractive index of the lens material

$R_1$  = the radius of curvature of the incident lens surface [m]

$R_2$  = the radius of curvature of the exit lens surface [m]

$d$  = the thickness of the lens measured at the optical axis

(7.16)

With a very thin lens, this equation simplifies into the *thin-lens equation* also called the *lensmakers equation*:

$$\frac{1}{f} = (n - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right) \quad (7.17)$$

In Figure 7.17 the imaging properties of a positive lens are shown<sup>4</sup>.

The object that is located at a distance  $S_1$  from the left side of the lens is imaged at a distance  $S_2$  from the right side. The rays really pass through the image, which means it is a real image. As will be shown in the next

<sup>4</sup>As of this example, the light rays are represented by lines instead of arrows and the light is assumed always to come in from the left and exit from the right, according to the sign convention. Further the optical axis is always represented by the centre line like in the previous figures.

figure, a real image can only be realised under the condition that  $S_1$  is larger than  $f$ . With thin lenses the location of the image can be approximately determined by using two rays of which the path or *trace* is followed. This process is called *ray tracing*. The first ray is drawn through the middle of the lens and does not change direction nor position. This is allowed as long as the lens is very thin. The middle of the lens will then behave like two parallel surfaces.

The second ray is drawn parallel to the optical axis and will be refracted to the focal point of the lens. The image is found at the intersection of the first and second ray.

The relation between  $S_1$  and  $S_2$  is determined by the focal distance  $f$ , following a simple relation that can be derived by trigonometry. This relation is called the **Gaussian lens formula**:

$$\frac{1}{S_1} + \frac{1}{S_2} = \frac{1}{f} \quad (7.18)$$

A simple sanity check proves this formula, as an object at infinity will result in an image in the focal point and vice versa. It also is imaginable that a distance of the object from the lens at twice the focal length results in full symmetry of object and image.

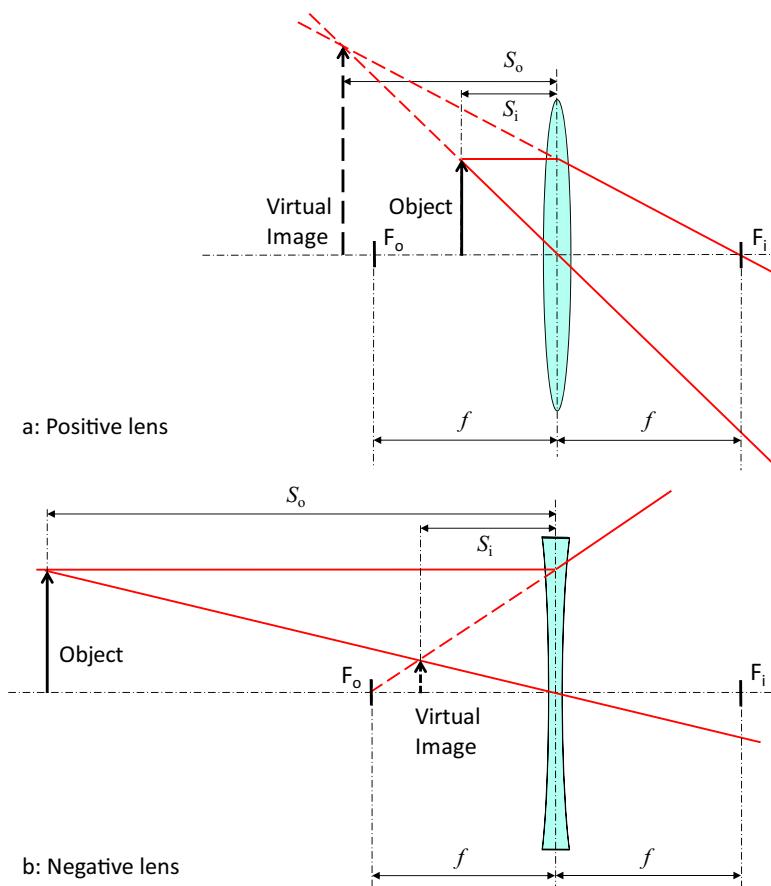
### 7.3.1.3 Magnification

In the previous example it was shown that the size of the image differs in most cases from the size of the object, resulting in a magnification  $M$  that depends on the position of the object relative to the focal point of the lens. This magnification can also be derived by means of straightforward trigonometry, showing that the magnification is proportional to the ratio between the distances. As a consequence of the sign conventions that were defined previously, the magnification  $M$  of a lens has to be written as follows:

$$M = -\frac{S_2}{S_1} \quad (7.19)$$

The minus sign represents the image location at the opposite side of the optical axis as the location of the object.

As a next step, it is interesting to see what happens when the object comes closer to the lens. At the image location this will result in an increasing distance to the right until the object is located at the focal point, resulting in an image at infinity. When the object approaches the lens even further,



**Figure 7.18:** A positive lens will give a virtual image, when the object is closer to the lens than the focal length. A negative lens always gives a virtual image, independent of the position of the object.

first the image jumps from  $+\infty$  to  $-\infty$  and then the image will approach the lens again, but now from left as a virtual image. This is illustrated in the upper drawing of Figure 7.18, where the resulting diverging set of rays at the right side of the lens seem to originate from the virtual image.

The rules for magnification and location of the object and image are still valid in this situation and because of the negative value of  $S_2$ , the sign of the magnification becomes positive. This corresponds with the position of the image at the same side of the optical axis as the object. Also the position of the image can be derived from the Gaussian Lens formula, where  $S_2$  becomes negative if  $S_1 < f$ .

Because of the virtual focal point, it is to be expected that a negative lens

will give a virtual image. This is indeed confirmed in the lower drawing of Figure 7.18, where the image is derived with the same rays, one parallel to the optical axis and one through the centre of the lens.

It is however not possible to create a real image with a negative lens by shifting the object closer to the lens. This can be concluded by examining the two rays in the figure, when the object would shift to the right. Even past the focal point, the rays will remain diverging at the right side of the lens, while the virtual image always remains closer to the lens than the object.

## 7.3.2 Aberrations

The quality of the imaging by optical systems is determined by many factors. Aside from the influence of the wave character of the light as will be presented in Section 7.4, these errors can consist of aberrations and stray light, leading to a reduction of contrast and sharpness.

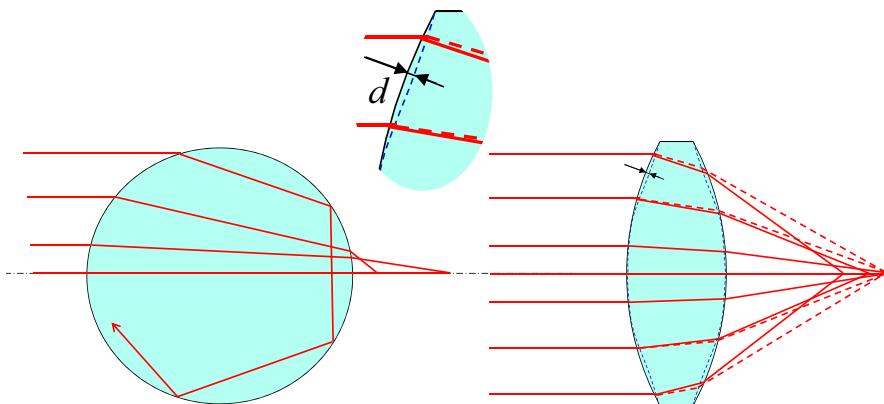
Stray light is caused by reflections on edges and the housing of the lens. Also insufficient polishing of the optical surface causes the occurrence of stray light, resulting in *flare* at the image.

All deviations of the trace of each ray in respect to the ideal trace to the image location are called aberrations. The three most typical aberrations, the *Spherical Aberration*, *Astigmatism* and *Coma* will be presented first, followed by the geometric and chromatic aberrations.

### 7.3.2.1 Spherical aberration

When presenting refraction at spherical surfaces, it was mentioned that parallel rays at different distances of the optical axis will refract to **almost the same focal point with only a small error**. This error is called the Spherical Aberration as it is caused by the pure spherical surface shape. Figure 7.19 illustrates this effect with the most extreme example of a positive lens, a full sphere.

In case of a purely spherical surface, the rays at a larger distance are refracted stronger than the rays more close to the optical axis. At a larger distance from the optical axis, the angle of incidence both at the entry surface and at the exit surface become increasingly large. The exiting point is closer to the optical axis where the normal to the surface is more horizontal, because of the refraction towards the optical axis. As a result, a stronger refraction occurs at the exit point of the lens than would be needed



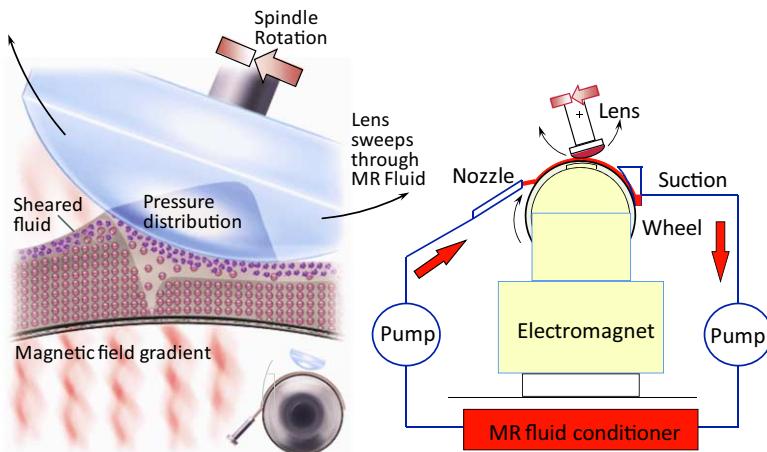
**Figure 7.19:** Spherical aberration at a lens with spherical surfaces. Rays at a different distance from the optical axis have a different focal length. The pure sphere at the left shows the most extreme example of this aberration, which makes a sphere unsuitable for normal imaging. With normal lenses spherical aberration can be solved by means of an aspheric departure  $d$  from the ideal sphere shape.

to reach the same focal point as where a beam closer to the optical axis is directed. Ultimately, with the most outer rays, the angles become so large that the angle of incidence at the exit surface becomes larger than the angle of total reflection and the light can not escape anymore.

Although the pure sphere is an extreme example, all lenses with spherical surfaces show spherical aberration, because the radius of curvature at a larger distance is too small to refract the outer rays to the same unique focal point. This reasoning implies that an increase of that radius as function of the distance can solve the problem. This deviation of the spherical surface is called *asphericity* with a *departure* value  $d$  as shown in the figure.

An aspheric surface can not be produced by the standard production methods for polished optical surfaces. Computer controlled milling, grinding and polishing machines are used to create these aspherical surfaces. An example of such a machine is the magneto-rheological polishing machine of the company QED shown in Figure 7.20. It uses the property of fluids that stiffen by a magnetic field to concentrate the polishing force very precisely on the surface of the lens where material has to be removed. The fluid is continuously refreshed and filtered, which means that the process parameters are not influenced by wear or floating glass particles.

Due to the high cost level of these machines the departure magnitude is a



**Figure 7.20:** Magneto-rheological polishing process used to create aspherical optical surfaces. The abrasive particles are suspended in a magneto-rheological fluid that stiffens in the magnetic field, applied near the optical surface to be polished. (courtesy of QED)

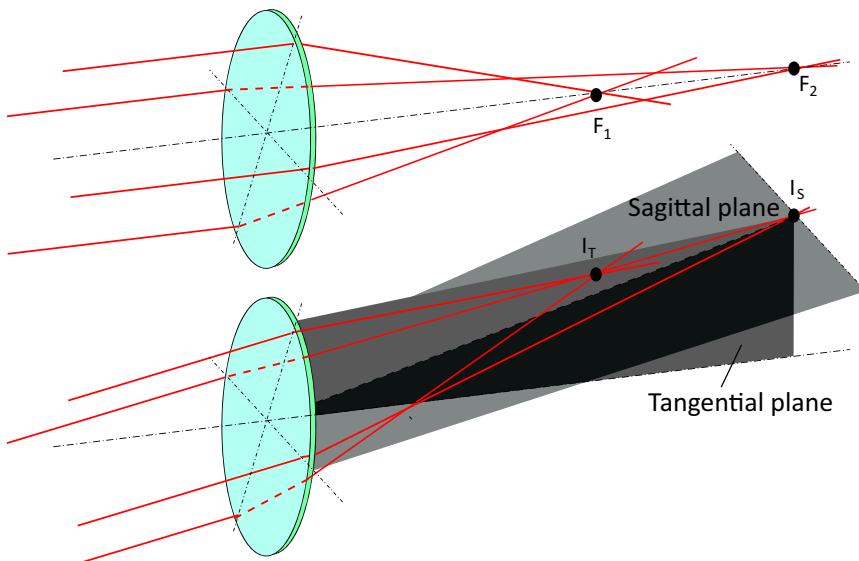
significant factor in the total cost of these optics.

### 7.3.2.2 Astigmatism

A second typical aberration is the astigmatism, as shown in Figure 7.21. This aberration is observed when rays that propagate in one plane through the optical axis have a different focal distance than rays propagating in an orthogonal plane through the optical axis.

This aberration occurs, when the radius of curvature of the lens surface is different for different planes through the lens. A lens surface having a combination of a spherical and a cylindrical profile is an example of a shape that causes spherical aberration. This causes the lens to become non rotation symmetric. With eye glasses, astigmatism is for this reason also called a *cylinder*. It is the most frequently observed aberration of the human eye.

The example shown in the drawing image of the figure illustrates that astigmatism can even occur for paraxial rays to an image location on the optical axis. With non-paraxial rays, the imaging of objects at some distance of the optical axis is analysed by tracing rays from different points on the object in two important planes as shown in the lower drawing of the figure. The first plane, the *tangential plane* which is also called the *meridional*

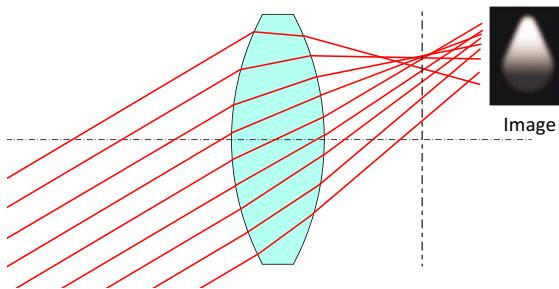


**Figure 7.21:** Astigmatism is the effect that is observed, when the rays in one plane have a different focal point than rays in a perpendicular plane. This is caused by a cylinder shape superimposed on the spherical shape of the lens. Astigmatism in the image of an object that is not located on the optical axis is recognised by different focal distance in two planes, the sagittal plane with image ( $I_S$ ) and tangential plane with image ( $I_T$ ).

*plane*, is defined by the point on the object and the optical axis. The second plane, the *sagittal plane*, is orthogonal to this plane and also includes the point on the object that is analysed. This means that the sagittal plane does not include the optical axis. With a single lens element it intersects the optical axis at the centre of the lens.

Rays in the tangential plane are called *tangential rays* and consequently rays in the sagittal plane are called *sagittal rays*. Astigmatism occurs when the sagittal and tangential rays have a different focal point and this can even occur with a rotation symmetric lens. In that case however the astigmatism for an object point on the optical axis would be zero.

A lens with a partly cylinder shape causes even astigmatism on the optical axis as was shown in the upper drawing of Figure 7.21 because of the non rotation symmetric shape.



**Figure 7.22:** Coma only occurs for non paraxial rays and is caused by a difference in focal length for rays that enter the lens at different positions. As a result the image looks like a “comet-tail”.

### 7.3.2.3 Coma

While spherical aberration and astigmatism can both occur even with paraxial rays and objects at the optical axis, other aberrations only occur with rays that correspond with an image that is not located on the optical axis. An example of this kind of aberrations is the coma as shown in Figure 7.22.

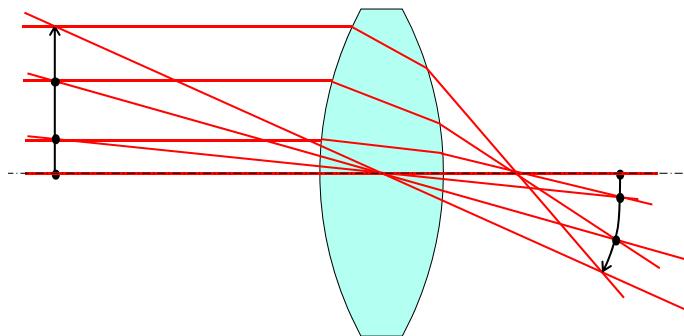
Coma is caused by the same phenomenon as spherical aberration. A larger angle of incidence at the entry surface of a lens results in an excessively refracted beam at the exit surface. This means that the rays that enter the lower part of the lens, where they are almost perpendicular to the surface, will be directed to a different focal point at a larger distance than the rays entering the upper part of the lens.

From the figure it can be seen that in this case also no real sharp focal point can be found. The image will look like a comet shape, depending of the position of the chosen image plane. Coma is recognised by a rather sharp and intense spot with a flare of decreasing intensity but increasing size in the direction of the optical axis.

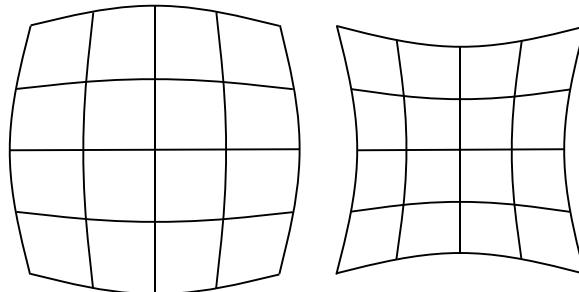
### 7.3.2.4 Geometric and chromatic aberrations

Other aberrations include geometric aberrations like *field curvature* and *distortion*, while *chromatic aberration* is observed when light with different wavelengths is imaged.

Field curvature means that a flat subject in the object plane is not imaged sharply flat at the image location but at a curved *focal plane*. This is shown in Figure 7.23. The object and image plane are often called a *field plane* to distinguish them from the *aperture plane* or the *pupil plane* that will be



**Figure 7.23:** Field curvature is the effect that the part of the object that is further away from the optical axis is imaged at another distance from the lens than the part that is close to the optical axis. Although again caused by the same mechanism as spherical aberration and coma it is not the same phenomenon.

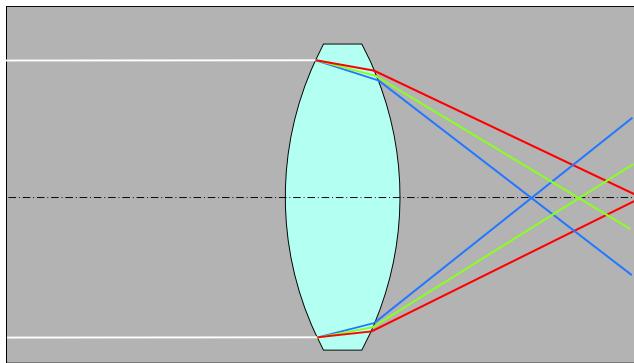


**Figure 7.24:** Barrel and Pincushion distortion are caused by a difference in magnification between parts of the object at a different distance from the optical axis.

defined later.

Distortion is an error in the magnification of a lens. When the magnification of the lens is depending on the distance of the object to the optical axis, the shape of the image of a large subject becomes distorted. Figure 7.24 shows the two most frequently observed distortions in for instance photo cameras, the *barrel* distortion where the magnification decreases for objects further away from the optical axis and the *Pincushion* distortion with the opposite effect. These two examples are only first-order linear types of distortion and in reality also higher-order distortion effects can occur.

Chromatic aberration is caused by the difference in refractive index of a transparent material for different wavelengths. This property is called



**Figure 7.25:** Chromatic aberration is caused by the wavelength dependent refractive index of the lens material. Different wavelengths (colours) will have a different focal length.

*chromatic dispersion* because it causes the well-known separation of colours in for instance a rainbow and a glass prism.

In Figure 7.25 an example with a *normal dispersion* is shown, where the refractive index increases with decreasing wavelength. This causes the focal point for blue light to be closer to the positive lens than the focal point for red light. Often this aberration also results in a wavelength dependent magnification, which is visible in photography as *purple fringing* at the outer parts of the image.

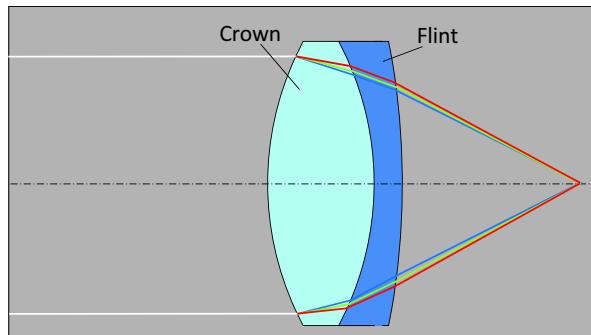
### 7.3.3 Combining multiple optical elements

The mentioned aberrations all have impact on the quality of the image and hence need to be reduced to an acceptable level, depending on the application.

By introducing an aspheric surface it was demonstrated how spherical aberration can be avoided and asphericity can also be used to reduce other errors like coma.

With low cost optics and in case of severe size limitations, like with the photo camera in a cell phone, this is the only way to solve these problems. Another example of such a limited optical system is the lens for the optical pick-up unit of a CD-player. These lenses are often made from plastic, molded in the desired aspheric shape.

In high performance imaging systems, this aspheric single-lens solution is not sufficient and a multitude of lens elements, both spherical and aspherical, are often combined to create an imaging system that is minimally



**Figure 7.26:** By combining a positive lens made of Crown glass and a negative lens made from Flint glass an achromatic doublet is obtained, where the chromatic aberration of a single lens element is compensated.

affected by the mentioned aberrations.

The first example of a multiple lens element is a combination of two lens elements, that are made of materials with different refractive properties, to correct chromatic aberration. When a positive lens, made of *Crown glass* is combined with a negative lens, made of *Flint glass* it results in the *achromatic doublet* of Figure 7.26.

Crown glass is a material with a very high refractive index and a corresponding high dispersion, while Flint glass has a low refractive index with a low dispersion. The resulting reduction of the overall dispersion comes at a price, because the focal length of the combination is larger than the focal length of the positive lens alone.

Although this combination gives an improvement, more measures are needed to correct a lens over a wide range of wavelengths. Nevertheless it is a building block that can be used in combination with more elements in advanced optical systems.

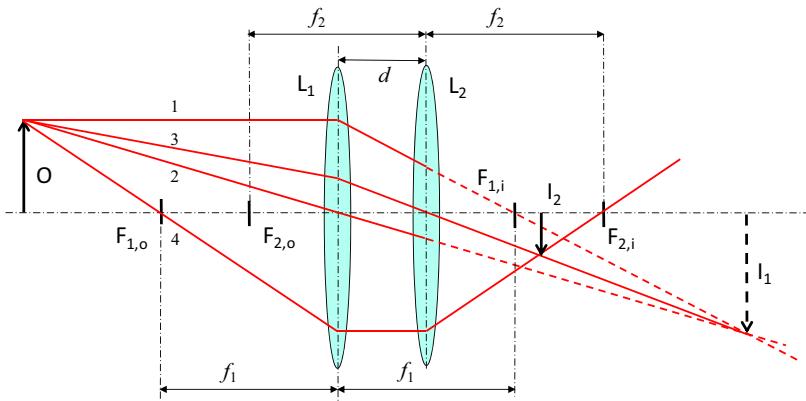
### 7.3.3.1 Combining two positive lenses

From the wide variety of combinations with positive and negative lenses, only the combination of two positive lenses will be explained in this section as representative example.

Figure 7.27 shows this configuration, where the distance between the positive lenses is smaller than the sum of their focal lengths.

The determination of the image location is done in two steps. First the image  $I_1$  is determined that would be created by the first lens  $L_1$  only.

As explained previously, this is achieved by drawing two rays. The first ray



**Figure 7.27:** Optical system, consisting of two positive lenses positioned close together. The image is determined by first determining the image by  $L_1$  via ray 1 and 2 and use that intermediate image to construct the image after  $L_2$  by means of an additional ray 3 drawn from  $I_1$  through the centre of  $L_2$  combined with ray 4 through both focal points.

is drawn parallel to the optical axis and will be refracted through the focal point at the image side  $F_{1,i}$ . The second ray propagates through the optical centre of  $L_1$  so it is not refracted. At the cross section of these rays, image  $I_1$  is drawn in a dashed shape, because in reality the rays will be intercepted by lens  $L_2$ .

With the help of image  $I_1$ , the third ray can be drawn, which is the one that would pass through the optical centre of  $L_2$  and image  $I_1$ .

As a last step to determine the image, the fourth ray propagates from the object through the focal point  $F_{1,o}$ , resulting in a refracted beam after  $L_1$  parallel to the optical axis, that will in its turn be refracted by  $L_2$  through its focal point  $F_{2,i}$ .

The image of the combination of the two lenses is found at the intersection of these last two rays.

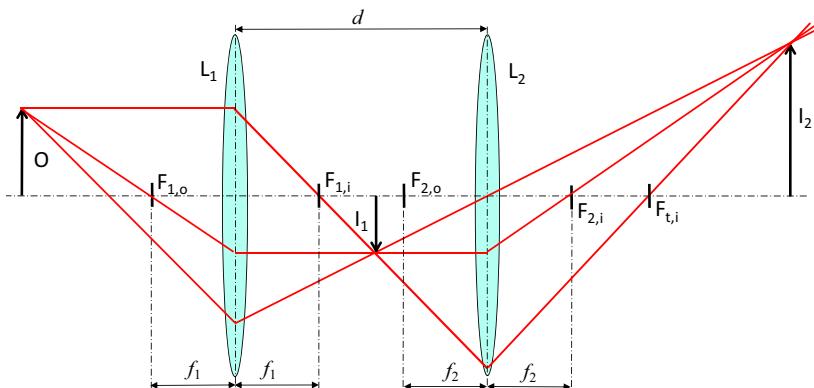
The two lenses appear to behave as one single lens and with a bit of trigonometry the following approximative relation for the focal length of the combination can be derived:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{d}{f_1 f_2} \quad (7.20)$$

Which for very small values of  $d$  relative to  $f_1$  and  $f_2$  becomes:

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} \quad (7.21)$$

It is also clear that the image is inverted, like with a single positive lens.



**Figure 7.28:** Optical system, consisting of two positive lenses positioned far apart. The image is also determined by first determining the intermediate image by  $L_1$ . The final image is determined straightforward, because the intermediate image is positioned between the focal points of both lenses.

A very different situation is shown in Figure 7.28, where the lenses are positioned further apart such that their distance is larger than the sum of their focal lengths.

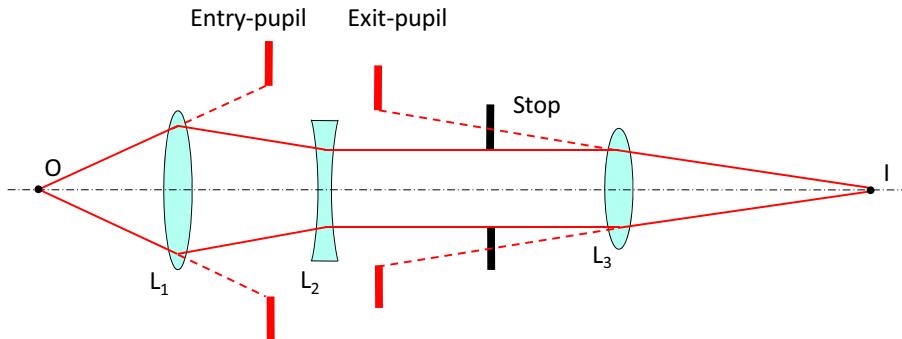
With this configuration, the image of the object by  $L_1$  is positioned between both lenses and is called the *intermediate image*. The second image is simply determined by taking the intermediate image as object for  $L_2$  and use the same rules as with a single lens system.

From the figure it appears that the inverting effect of  $L_1$  is reversed by  $L_2$ , resulting in an image with the same orientation as the object.

By tracing the ray from the object, that was parallel to the optical axis, the focal point of the combination  $F_{t,i}$  is found. When  $L_2$  is positioned closer to  $L_1$  in Figure 7.28, it appears that  $F_{t,i}$  will move more to the right. This can be concluded from the position of the image when the object is kept at a fixed distance from  $L_1$ . The three rays that converge at  $I_2$  will rapidly diverge. As a result the magnification of the combined two lenses is increased and the focal point is shifted. This effect indicates that with two positive lens elements a *zoom lens* can be created.

A special situation occurs when  $L_2$  is shifted so close to  $L_1$  that  $F_{2,o}$  becomes located at the same position as  $I_1$ . In that situation the second image  $I_2$  will be located at infinity as  $I_1$  is at the focal point of  $L_2$ .

When  $L_2$  is moved even further towards  $L_1$ ,  $I_1$  will enter the area between  $F_{2,o}$  and  $L_2$  resulting in a virtual image at the left side of the system, similar



**Figure 7.29:** An optical system with an aperture stop that limits the outer rays in the system. The entry-pupil is the image of the aperture stop observed at the entry of the system, while the exit-pupil equals the image of the aperture stop as observed at the exit of the optical system.

to the effect that was demonstrated a few pages back in Figure ???. It is true that at the very moment of crossing the focal point of  $L_1$ , the image swept from positive infinity to negative infinity and will again come closer from the left, when  $L_2$  is moved even further.

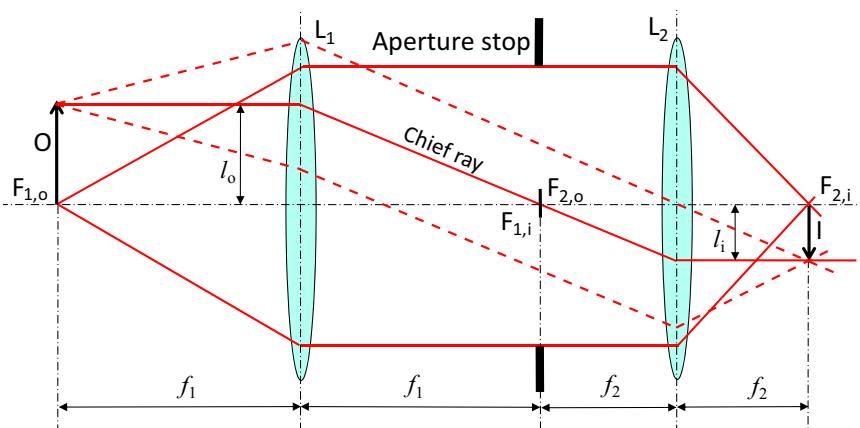
Ultimately both focal point will coincide which results in the double-telecentric lens that will be presented after the introduction of the important aperture stop and pupil.

### 7.3.4 Aperture stop and pupil

All optical imaging systems have one important limitation that is caused by their size. In Section 7.4 on physical optics it will be shown, how the ultimate imaging properties of an optical system are determined by the maximum capture angle of the rays that can pass the system. This quality of capturing as much as possible light is also directly related to the law of conservation of radiance, as introduced in Section 7.1.3.

A geometrical measure for this dimensional property is the *pupil* that defines which rays are captured and transmitted by the optical system and which rays are blocked.

With a single lens element, this pupil is determined by the diameter of the element itself. With a more complicated system, consisting of a multitude of lens elements, the outer rays are in principle limited in a not well defined manner, depending on the direction of the rays and the position and sizes of the different elements.



**Figure 7.30:** When combining two positive lenses with coinciding focal points, a double-telecentric optical system is obtained, where an object at  $F_{1,o}$  will be imaged at  $F_{2,i}$  with a magnification equal to the ratio of the focal lengths. The chief ray through the joint focal centre is parallel to the optical axis both at the object and at the image site. The position of the object or the image sensor has no influence on the magnification. A symmetric cone of light from the object around the chief ray, indicated by the dashed rays, will also create a symmetric cone of light at the image.

It is not preferable to determine the pupil by the mountings of the lens elements by means of metal constructions with either springs, screws and/or glue, because this would lead to scattering of light in undefined directions. To solve these issues, preferably a separate element is added that determines this pupil. This element is called an *aperture stop*, often just shortly named “the stop”. Other names are *iris diaphragm* or just *iris*.

Figure 7.29 shows a schematic example of an optical system that consists of three lens elements, where an aperture stop is inserted between lens  $L_2$  and  $L_3$ . The first observation that can be made is that the most outer rays that pass through the system are no longer determined by the size of the lenses but only by the aperture stop. The second observation is that only the image of the stop by the lenses between the stop and the observer can be seen from outside. This image is called the *entry-pupil* at the object side and the *exit-pupil* at the image side.

### 7.3.5 Telecentricity

When the focal points of two positive lenses coincide, the configuration obtains a special property, called *double-telecentricity* and is explained with the dual lens-element system of Figure 7.30.

A paraxial ray from **O** will be refracted by **L<sub>1</sub>** through the joint focal point and continue its path towards **L<sub>2</sub>**, where it will be refracted parallel to the optical axis again.

The location of the image will be at **F<sub>2,i</sub>**, because all rays of any point of the object will be refracted into a set of parallel rays in the space between the lenses. As a consequence they themselves will be refracted to the focal point of the second lens element **L<sub>2</sub>**.

The aperture stop in a telecentric system is located at the joint focal point. The ray through the joint focal point is called the *chief ray*. Officially a chief ray is defined as any ray under an angle with the optical axis that passes through the centre of the aperture stop.

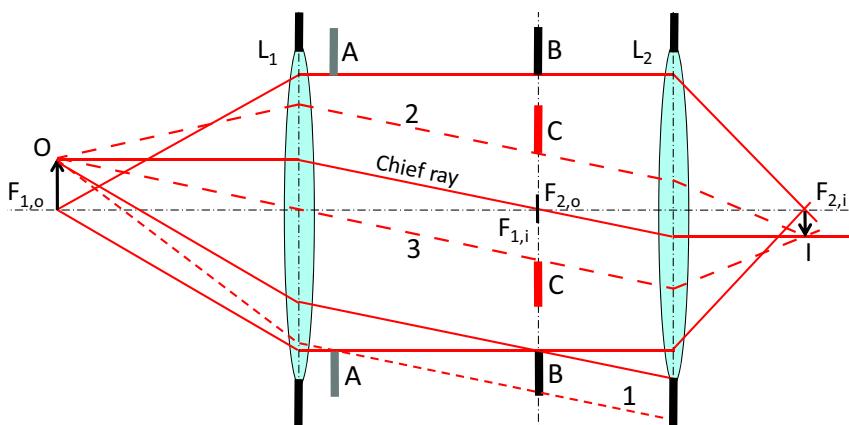
With a double-telecentric system a paraxial ray from any point on the object will be a chief ray and become also a paraxial ray at the image side of the system. As a consequence, the magnification of the system is only determined by the ratio of the focal lengths:

$$M = -\frac{l_i}{l_o} = -\frac{f_2}{f_1} \quad (7.22)$$

This means that a displacement of the object in any direction left or right from the drawn position in Figure 7.30 will not change the size of the image, as the path of the drawn paraxial ray will not be influenced by this displacement.

It should be noted that even with this constant magnification, the focal position of the image will change with a displacement of the object along the optical axis. This means that these kind of lenses are specially chosen when the requirements on magnification are more severe than the requirements on focal distance. This is the case with for instance the projection lens of the wafer stepper that was introduced in Chapter 1. When the wafer with photo-sensitive resist is placed at the image location with a slight error in its distance to the lens, this error will only result in a less than perfectly sharp image but not in a change of the dimensions.

In the described configuration, this magnification is maintained both for a position change of the image sensor and for a position change of the object, which makes it a double-telecentric optical system. Also examples with single-telecentricity exist as will be presented later.



**Figure 7.31:** Different positions of the aperture stop have different effects. Locating the aperture stop at **A** instead of the preferred position **B** at the joint focal point will result in ray 1 to be blocked at an undefined location. A smaller aperture stop at **C** clearly defines the cone of light bounded by ray 2 and 3.

### 7.3.5.1 Pupil in a telecentric system

Figure 7.31 shows a double-telecentric lens configuration with the aperture stop at different locations to illustrate the different effects.

At first sight it seems not important where the aperture stop is exactly located, because the rays between the lenses are propagating in parallel. Still it is preferred to position this stop at the coinciding focal points, as there the impact of the size of the object and image is minimised.

When the stop would be located near lens L<sub>1</sub> as shown in the figure at point **A**, a ray coming from the top of the object going just past the lower part of the stop, would end-up outside lens L<sub>2</sub>. This means that the surroundings of L<sub>2</sub> would determine another undefined aperture stop for these lower rays. It can be concluded that the maximum allowable diameter of the stop is determined both by the size of the lens elements, the size of the object and/or image and by the distance of the aperture stop to the joint focal point.

It is also shown in the figure that the aperture stop defines a cone of light around the chief ray. With the double-telecentric lens this cone is always pointing in a parallel direction to the optical axis.

The image of the stop of a double-telecentric lens is at infinity because of the position at the joint focal point. In practice this has given cause to the name *pupil plane* for the orthogonal plane to the optical axis that is located at the aperture stop.

Also with non-telecentric optics it is preferred that the rays from any spot on the object propagate parallel at the plane where the aperture stop is located. This is the case with all optics that need a controllable aperture stop, like photographic camera lenses. This can be seen from Figure 7.31 where for instance a smaller aperture stop at position **C** will only influence the amount of light. As each point on the object will still be imaged by means of light rays through the smaller aperture the overall image shape is not altered by shadowing a part of the image. For that reason also with non-telecentric lenses this plane where the aperture is located is called a pupil plane.

With the physical optics of the following section, this pupil plane will further prove to be a very important area to determine the quality of an imaging system.

### 7.3.5.2 Practical applications and constraints

The shown example with only two positive lenses represents an almost ideal situation. When a different amount of elements are used like in Figure 7.29, there is not always a joint focal point defined. In those cases still single-telecentricity can be created, depending on the location of the aperture stop inside the lens system.

The first possibility is an *object-space telecentric* system, where the aperture stop is located at the first focal point inside the system when entering from the object side. In that case the entry-pupil is located at infinity.

An example of such an object-space telecentric lens is the measurement microscope from Chapter 3. A small displacement in the vertical direction is not allowed to cause an error in the measurement of the object size and object-space telecentricity prevents this error.

The second possibility is an *image-space telecentric* system, where the aperture stop is located at the first focal point inside the system when entering from the image side. This causes the exit-pupil to be located at infinity.

Image-space telecentric lenses are preferred in digital photography cameras. Non-telecentricity at the image sensor has as consequence, that the cone of light to the image is not symmetric around a paraxial ray but tilted under an angle with the optical axis. This can lead to image artifacts, due to the stack of a colour filter and multi-lens array before the image sensor. Only an image-space telecentric lens guarantees that for instance no red light ends up at a blue sensor element.

Due to size constraints, practical photographic lenses can only approximate

telecentricity. Perfect telecentricity would require that at any position on the image sensor a full symmetric cone of light could be delivered by the objective. With the shown example of Figure 7.31, this requirement would imply that  $L_2$  would need to be much larger than the image sensor and this size would certainly not fit in a normal camera with a full frame sensor of  $24 \times 36$  mm. As a result *vignetting* would occur, a darkening of the outer part of the image. With a smaller sensor like the APS-C type of  $23 \times 15$  mm, the telecentricity is more easily to approach without increasing vignetting. The telecentricity of a camera lens can be checked by estimating the distance of the exit-pupil when looking at the image side through the lens. It should be as far away as possible but will in most cases not be located much further away than half-way the lens.

## 7.4 Physical Optics

One of the models to describe the properties of light was shown to be based on the theory of electromagnetic wave propagation.

In Chapter 2 it was explained that electromagnetic waves consist of a combination of two oscillating orthogonal vector fields, an electric **E** and a magnetic **B** field. In Chapter 5, the interrelation between magnetic and electric fields was defined by Faraday's and Ampère's law of the Maxwell equations and these laws are used in dedicated modelling software for the calculation of the real properties of light in optical systems. In this section the wave character of the light is used to explain the performance of optical systems without extensive calculations. Instead, graphical representations will be used of only one of the two interrelated vector fields. The electric **E**-field is often chosen for this purpose and is presented graphical as a sine wave.

An unfortunate fact that is encountered in analysing the physical properties of optics, is caused by the extreme frequency of the related electromagnetic waves. Electric or magnetic oscillating fields with frequencies in the order of  $10^{15}$  Hz are impossible to measure directly. As mentioned before, only the average optical irradiance  $I_r$  can be detected by means of photo detectors that convert the related flow of photons into an electric current.

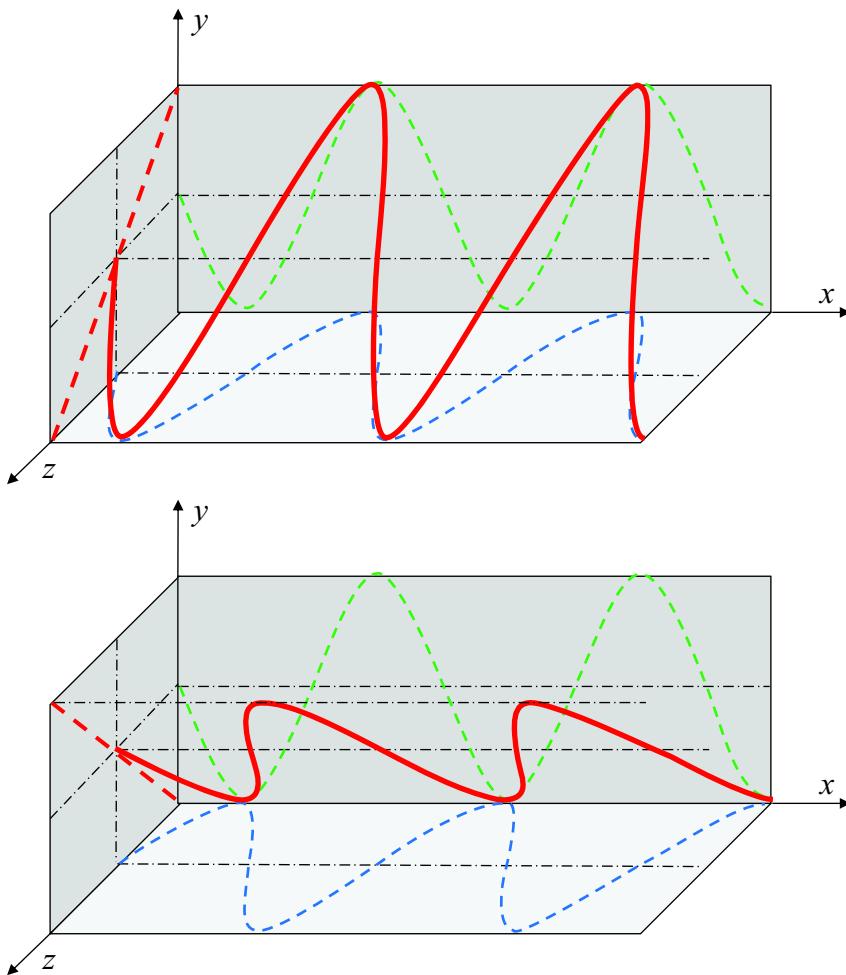
The irradiance of an electromagnetic wave is proportional to the power of the sinusoidal electric field, given by the following expression:

$$I_r = c n \epsilon_0 \mathbf{E}_{\text{rms}}^2 = \frac{c n \epsilon_0}{2} \hat{\mathbf{E}}^2 \quad [\text{W/m}^2] \quad (7.23)$$

In the following subsections, first the concept of polarisation is introduced, because this concept is widely used in optical measurement systems. This is followed by the explanation of the interference effects of the interaction between different waves and the theory on diffraction and gratings. The section will be completed by deriving the diffraction limited resolution of an optical system, where the influence of the capture angle on the quality of the image will be demonstrated.

### 7.4.1 Polarisation

The *polarisation* of an electromagnetic field is based on its transversal character and defines in which direction the electric field is oscillating. To see what this means, the electric field of the light is represented in the 3D vectorial graph of Figure 7.32. It applies an orthogonal coordinate system,



**Figure 7.32:** The electric field of linear polarised light, represented in an orthogonal coordinate system, defined by  $x$  in the direction of propagation of the wave and  $y,z$  under  $45^\circ$  with the polarisation direction of the electric field. The projection on the  $y - z$  plane is a line. in the upper part the  $y$  and  $z$  components are in phase while in the lower part they are  $180^\circ$  out of phase. Shifting the phase of one of the components with  $180^\circ$  results in a rotation of the polarisation direction of  $90^\circ$  on the  $y - z$  plane.

with  $x$  being the direction of propagation.

In this example, the field oscillates in one direction and the coordinate system is chosen such that the oscillation direction is under  $45^\circ$  with the  $y$  and  $z$  axis.

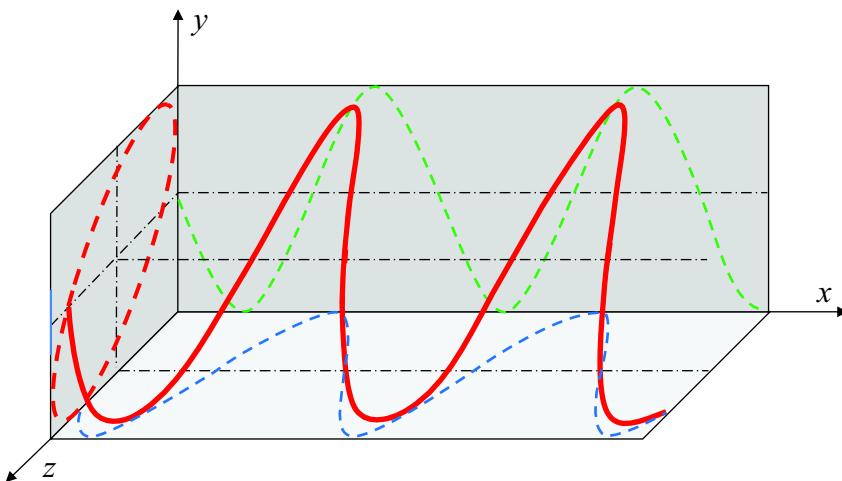
The components of the wave in the  $x - y$ ,  $y - z$  and  $x - z$  plane are deduced by simple projection. Projection of the wave in the  $y - z$  plane gives a straight line, for which reason this oscillating electric field is called *linear polarised*. The amplitude of each component in the  $x - y$  and  $x - z$  plane depends on the polarisation direction in respect to the chosen coordinate system, which in this example results in an equal amplitude.

The polarisation direction of the light emitted from a light source is determined by its origin and generally normal light sources like a light bulb or the sun are “randomly polarised”. This means that there is hardly any phase relationship. In order to create a linearly polarised light source, a *polariser* can be used, an optical element that only transmits the polarisation component of the light that corresponds to the orientation of the polariser. Polarisation is created in nature at reflecting surfaces, because the reflection of an electromagnetic wave is depending on the polarisation direction relative to the orientation of the reflecting surface. Light polarised in the plane of incidence, defined by the normal and the propagation direction, is called p-polarised light and reflects less than light polarised in the orthogonal direction, the s-polarised light. This effect is maximum at the *Brewster's angle*  $\vartheta_B = \arctan(n_2/n_1)$ , which for glass amounts to  $\approx 56^\circ$ . For this reason sunglasses with polarising glass reduce the effect of reflected light from the sea or other reflecting surfaces that has become polarised by the reflection.

A laser can be made to emit only one polarisation direction by inserting a quartz plate at the brewster angle inside the resonating chamber. The stimulated excitation will result in photons that behave in the frequency domain as having almost the same frequency, phase and polarisation as the photon that caused the excitation. Especially a well designed Helium Neon laser can create an almost ideal polarisation that can be used in position measurement systems as will be presented in the Chapter 8.

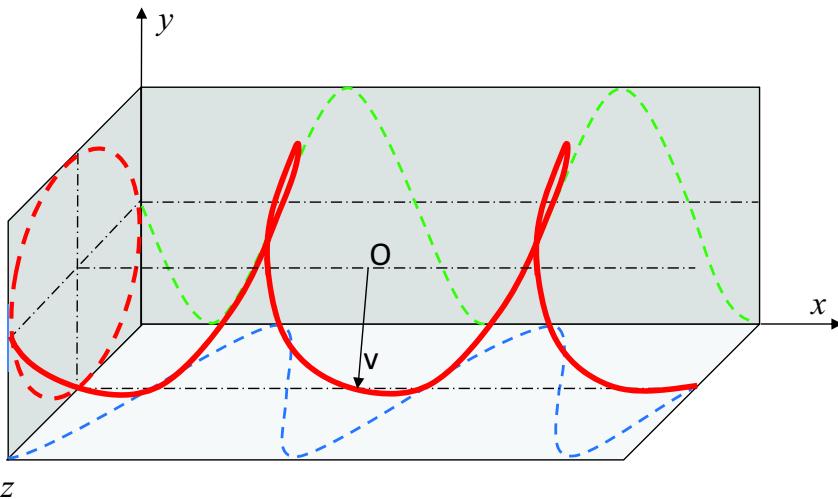
#### 7.4.1.1 Birefringence

It is often observed that the propagation speed of light inside a transparent material depends on the polarisation direction. This optical anisotropy is called *birefringence* and is caused by the atomic structure of the material. Especially crystalline materials show this property. Birefringent materials can be used to change the polarisation direction of light by inserting a certain length of this material in the beam. Starting with linear polarised light the anisotropic directions of this birefringent material need to be orientated under  $45^\circ$  with the polarisation direction of the light according to the coordi-



**Figure 7.33:** Elliptical polarised light. When the  $y$  component relative to the  $z$  component of the electric field of an electromagnetic wave is slightly different from  $180$  or  $0^\circ$  the projection of the electric field in the  $y - z$  plane becomes an ellipse.

nate system from Figure 7.32. As a result both polarisation components of the wave will no longer have the same phase relationship because one of them is delayed in respect to the other. What this means is illustrated in Figure 7.33 where the phase difference between both components is slightly less than  $180^\circ$ . As a result the projection of the combined wave on the  $y - z$  plane will no longer be a straight line but elliptical. For this reason this electric field is called *elliptical polarised*. The ellipticity is depending on the phase difference and in case the phase difference is  $\pm 90^\circ$  the projection on the  $y - z$  plane becomes a circle, as shown in Figure 7.34. This *circular polarised* light will give an equal amplitude of the projection in any direction independent of the chosen coordinate system. One can visualise the propagation of a circular polarised vector field as a rotating vector  $\mathbf{v}$  originating at a line  $\mathbf{O}$  in the propagation direction pointing to a perfectly symmetrical spiral around the line of origin. The rotation direction of the spiraling vector depends on the sign of the phase difference which means that two different versions of elliptical or circular polarised light exist.



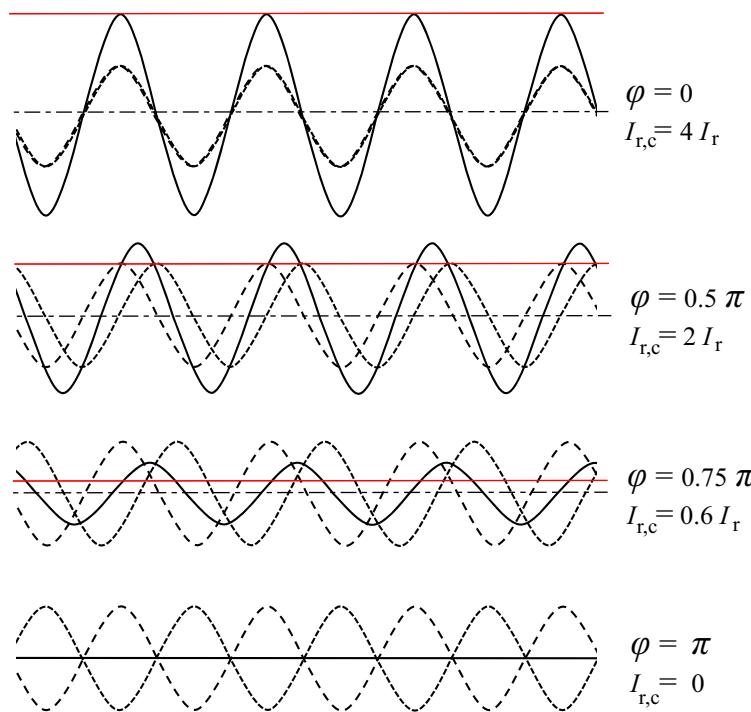
**Figure 7.34:** Circular polarised light. When the phase relationship between the  $y$  and  $z$  component of the electric field of an electromagnetic wave equals  $90^\circ$ , the projection of the electric field in the  $y - z$  plane becomes a circle. The spiral can rotate left or right handed depending whether the phase difference is positive or negative.

Polarisation as such is more complicated than explained in this summary but this basic understanding is sufficient to model optical systems for engineering purposes. In Chapter 8 the manipulation of the polarisation direction by using thin plates of birefringent material will be applied to enable the realisation of very accurate incremental optical measurement systems.

## 7.4.2 Interference

In Chapter 2 it was shown that the momentary values of different fields at a certain location can be added together. When two beams of light with an equal wavelength are combined, the resulting combination will show *interference* effects. These effects are determined by the electric field amplitude  $\hat{\mathbf{E}}$  and phase difference  $\varphi$  of both beams of light.

For two beams with an equal field amplitude, the effect of the phase difference is expressed mathematically as follows: When the electric field magnitude of both beams is given by  $\mathbf{E}_1 = \hat{\mathbf{E}} \sin(\omega t)$  and  $\mathbf{E}_2 = \hat{\mathbf{E}} \sin(\omega t + \varphi)$ , the



**Figure 7.35:** Interference between two waves with the same wavelength and amplitude but different phase. The resulting electric field amplitude can range between zero and twice the amplitude of each wave giving a combined irradiance  $I_{r,c}$  of maximum four times the irradiance  $I_r$  of each wave.

combined electric field becomes:

$$\begin{aligned} \mathbf{E}_c &= \hat{\mathbf{E}} (\sin(\omega t) + \sin(\omega t + \varphi)) = 2\hat{\mathbf{E}} \left( \sin\left(\frac{2\omega t + \varphi}{2}\right) \cos\left(\frac{-\varphi}{2}\right) \right) \\ &= 2\hat{\mathbf{E}} \cos\left(\frac{\varphi}{2}\right) \sin\left(\omega t + \frac{\varphi}{2}\right) \end{aligned} \quad (7.24)$$

First of all it can be concluded that the phase of the resulting wave is the average of the phase of both waves.

Secondly the amplitude becomes twice the value of the amplitude of one beam, when  $\varphi = 0$ , and zero, when  $\varphi = \pi$  or  $180^\circ$ . This is shown in Figure 7.35, where the upper graph shows the situation, when  $\varphi = 0$ . This situation where both amplitudes add to a double amplitude is called *constructive interference*. The lower graph shows the effect of *destructive interference* with  $\varphi = 180^\circ$ , where both amplitudes cancel each other.

When measuring the resulting electric field, only the irradiance can be observed. The irradiance can be determined by using the amplitude term of Equation 7.24 in Equation (7.23):

$$I_{r,c} = \frac{cn\epsilon_0}{2} \hat{\mathbf{E}}_c^2 = \frac{cn\epsilon_0}{2} \left(2\hat{\mathbf{E}} \cos\left(\frac{\varphi}{2}\right)\right)^2 = 2cn\epsilon_0 \hat{\mathbf{E}}^2 \cos^2\left(\frac{\varphi}{2}\right) \quad (7.25)$$

After expanding the squared cosine term with its trigonometric identity, this equation becomes:

$$I_{r,c} = 2cn\epsilon_0 \hat{\mathbf{E}}^2 \left(\frac{1 + \cos\varphi}{2}\right) = cn\epsilon_0 \hat{\mathbf{E}}^2 (1 + \cos\varphi) \quad (7.26)$$

The resulting irradiance of the combined electric field varies between  $2cn\epsilon_0 \hat{\mathbf{E}}^2$ , when  $\varphi = 0 + n \cdot 2\pi$ , and zero, when  $\varphi = \pi + n \cdot 2\pi$ , with the integer  $n \geq 1$ .

The irradiance of the separate fields is equal to:

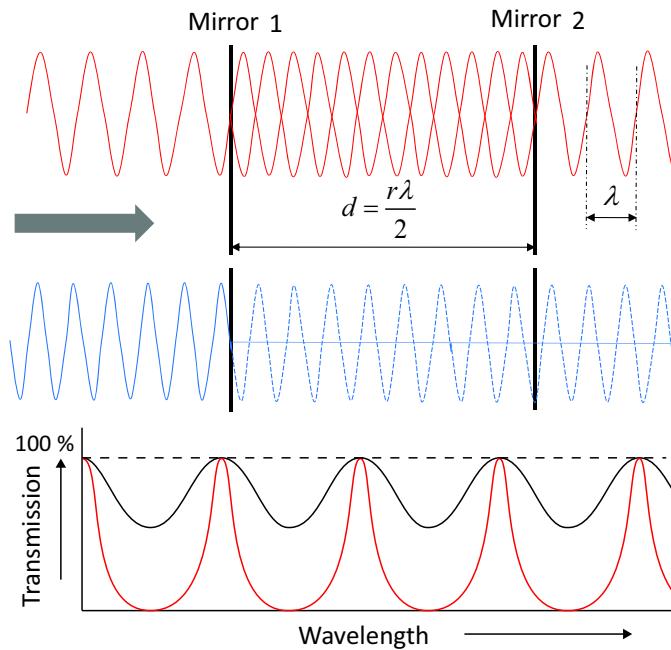
$$I_{r,1} = I_{r,2} = I_r = \frac{cn\epsilon_0 \hat{\mathbf{E}}^2}{2} \quad (7.27)$$

This irradiance level is a factor four below the irradiance level of the combined fields at full constructive interference. This is caused by the squared relation between the double amplitude of the combined fields and the irradiance. Although this result seems to be contradicting to the law of conservation of energy, in reality an optical system that shows constructive interference, also shows an equal amount of destructive interference at another location. The energy from the destructive interference area will add to the energy at the other area and the total energy within the system is not changed. This effect will be mentioned at different occasions in the following pages.

#### 7.4.2.1 Fabry-Perot interferometer

Interference is frequently used in optical measurement systems to determine distances and shapes of objects. As an example two beams of light that originate from the same coherent light source are guided over two different paths, one fixed and one changing, to one sensor location where they interfere with each other. The irradiance of the re-combined beams becomes determined by their path difference, which induces a timing or phase difference. By measuring the combined irradiance, the change in the length of the second path can be determined.

This measurement principle is called an *interferometer*. Many versions of an interferometer exist and in the next chapter the Michelson interferometer



**Figure 7.36:** Fabry-Perot interferometer consisting of two parallel mirrors at a fixed distance. The upper graph shows the situation where the distance of the mirrors equals an integer amount of half the wavelength of the light resulting in a high transmission because of optical resonance. In the middle graph part the wavelength does not “fit” between the mirrors so the transmission is zero. the lower graph shows the transmission as function of the wavelength where the black trace corresponds with a low finesse and the red trace with a higher finesse.

will be introduced to measure distances. Here the *Fabry-Perot interferometer* is further investigated as an example as this principle was previously presented as the optical resonator for increasing the coherence of a laser.

Figure 7.36 shows the principle of a Fabry-Perot interferometer. Two mirrors are positioned perfectly parallel at a fixed distance. Mirrors have always some transmission. A full 100 % reflection is not possible so some photons will pass the first mirror and enter the *cavity* between the mirrors, where most of them will be reflected at the second mirror back to the first mirror, reflected again and so on.

Due to their wave behaviour something special occurs, when the distance  $d$  of the mirrors equals an integer ( $r$ ) amount of half the wavelength of the light ( $d = r\lambda/2$ ). In that case the reflected light at the left entry-mirror will

show constructive interference with the light just entering the cavity and as a result a standing wave occurs.

This standing wave manifests itself as if the interferometer becomes fully transparent for that wavelength as shown in the upper half of the figure. Other frequencies, that do not fulfil the requirement  $\lambda = 2d/r$ , will result in less constructive or even destructive interference and can not pass the interferometer, depending on the amount of destructive interference. This system acts like an optical notch filter.

The capability of this filter to distinguish different wavelengths is called *finesse* and is determined by the ratio between reflection and transmission of the mirrors of which the interferometer is made. When the transmission is large also many other frequencies will pass, because there is hardly any internal reflection, which would cause destructive interference. This is for instance the case with glass in a normal window. On the other hand, the more the mirrors become ideal reflectors the less other frequencies will pass.

A special version of a Fabry-Perot interferometer is the *dichroic coating* that is used to prevent reflections on a lens or increase reflections on a mirror. It consists of many thin layers of material with a different refractive index that each determine a very small cavity tuned to a specific frequency. By combining a multitude of layers it is possible to either create an optical filter with a very high finesse or a more wide-band behaviour like with photographic lenses.

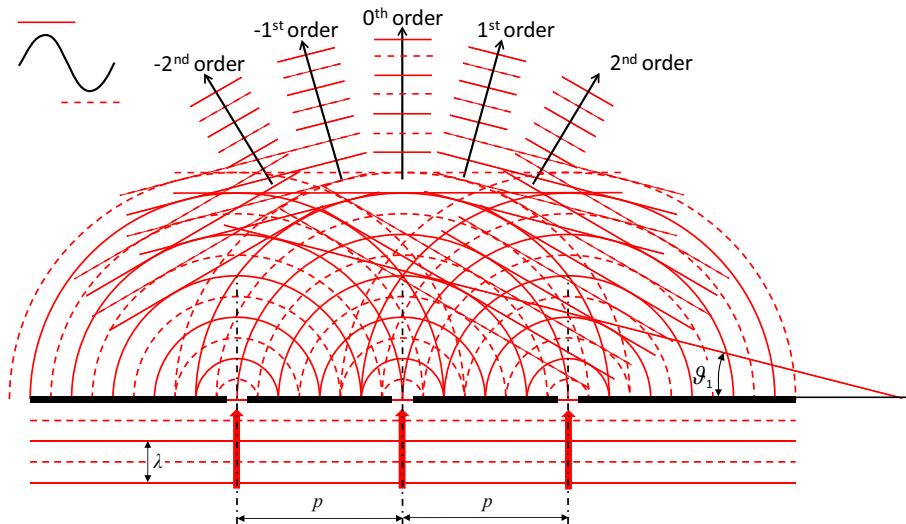
### 7.4.3 Diffraction

The famous Dutch scientist Christiaan Huygens (1929 – 1695) postulated in his work *Traité de la lumière*, that wave propagation takes place because every point of incidence of a wave will act as a source of a new spherical wave with the same frequency and in phase as the original wave.

With a multitude of these kind of sources he could construct the resulting wavefront of light by connecting the wavefronts of each individual source, while having by definition the same phase.

A flat wavefront would then consist of an infinite amount of separate wavefronts, where only the parallel plane to the original wavefront would satisfy this phase relationship.

His theory was later rejected for reason of not sufficient validity in all cases and the particle based theory with photons was found more suitable. Nevertheless the theory of Huygens was not so wrong and it can be used to model the effect of diffraction of light at a grating in a graphical way.



**Figure 7.37:** Diffraction of light at a transmissive amplitude grating. The holes or slits behave like separate light sources with a fixed phase relationship, determined by the incoming wavefront. By drawing lines tangential to the circles with equal phase of all slits, the wavefront of a diffraction order is found. In the intermediate directions, wavefronts with an opposite sign result in destructive interference. The angles of the resulting diffraction orders depend on the ratio between the distance of the holes ( $p$ ) and the wavelength of the light. Only for the  $2^{\text{nd}}$  order all tangential lines, connecting the wavefronts of the three holes, are shown.

#### 7.4.3.1 Amplitude gratings

A first example of a diffraction grating is the transmissive *amplitude grating* as shown in Figure 7.37. The name is based on the property of this grating that light from an area with a high electric field amplitude is combined with light from a low electric field amplitude. The grating in this example consists of three very small parallel slits that are defined orthogonal to the plane of sight in a further completely opaque plate.

A flat wavefront of light with only one wavelength is drawn parallel to the grating, approaching this grating from below<sup>5</sup>.

According to Huygens this incoming light will create separate sources of

<sup>5</sup>Always remember that a wavefront drawing is a snapshot of a normally travelling wave of the electromagnetic field of light. It is used in all examples of diffraction only to show the phase relationship between the different light waves

light at the slits with equal amplitude and phase. Orthogonal to the plane of sight these sources create a line source so the resulting wavefront will be like a cylinder of which only the cross section is shown in the figure.

The resulting wavefronts are found by connecting the wavefronts of the different line-sources that are mutually in phase. Several favourable versions of these tangential lines can be found as is shown in the figure. The orthogonal directions to each of these tangential lines are directions where the light from all sources interfere in a constructive way and light will be observed in those directions. The other directions in between show destructive interference as both light with a positive and a negative phase is combined. As a result, the diffraction at this grating manifests itself as a series of separate light beams under an angle  $\vartheta$  that is determined by the wavelength of the light and the distance of the slits in the grating. These separate beams are called diffraction orders with increasing number when starting from the optical axis. In principle the diffraction is mainly observed at some distance from the slits, the *far field*. At a closer distance the amplitude of the field of each slit at the location of observation will be different and at a distance in the order of the period of the grating, the light of only one slit will be observed to be propagating in all directions.

From the figure it can be concluded that the angle between the orders will be larger with a larger wavelength and/or a shorter periodic distance of the slits. This relation can be derived by trigonometry which results in:

$$\vartheta_N = \arcsin \frac{N\lambda}{p} \quad (7.28)$$

where  $\lambda$  is the wavelength of the light,  $p$  is the periodicity of the grating (distance of the centres of the slits) and  $N$  is an integer, equal to the order number.  $N$  can be positive and negative, which means that the diffraction occurs at two sides. It can be noted that the maximum amount of orders equals the ratio between  $p$  and  $\lambda$  as the value of a sine is limited to one. When using light of different wavelengths, diffraction would result in spatial separation of the different wavelengths, giving the colouring effects with visible light as for instance can be observed when looking to a CD-disk.

It is also true, that a larger irregular grating that consists of a number of sub-gratings with different periodic distances would result in spatial separation of the diffracted parts that belong to the different sub-gratings, even with one wavelength of light. The shorter periodic distances will have a larger diffraction angle than larger distances. A shorter periodic distance corresponds with a higher spatial frequency of the grating lines. The fact

that diffraction distinguishes areas with different spatial frequencies will be used later when analysing the imaging quality of an optical system.

The optical effect of the transmissive amplitude grating with transparent slits in an opaque plate is also observed with a reflective amplitude grating. In that case the light approaches the grating from above, while the slits are replaced by reflecting stripes on a non reflecting surface. The reflective stripes will each act as a source of light like the slits of a transmission grating and the resulting diffraction orders can be determined in an identical way.

#### 7.4.3.2 Phase gratings

Amplitude gratings have one disadvantage, as they reduce the total amount of light by the opaque or absorbing part of the surface. This property limits the maximum reflectivity or transmissivity to approximately 50 %.

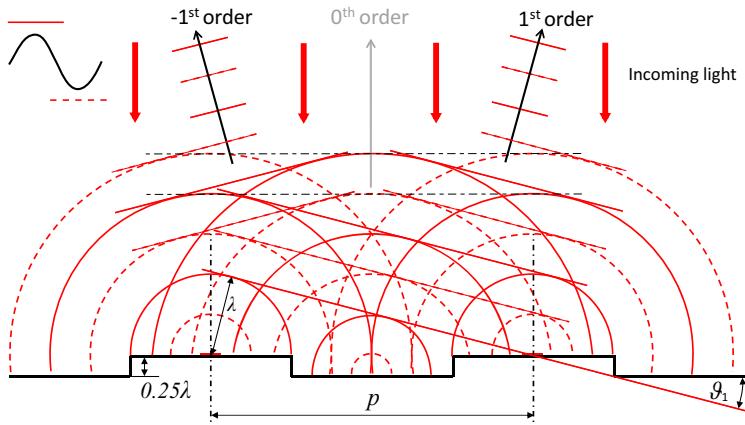
One might decrease the loss of light by making the slit wider than the opaque part, but then the interference will be less. Ultimately the object becomes just transparent or reflective, which was not the aim of the grating.

Fortunately the 50 % loss of light can be avoided by using a *phase grating*. In its reflective form, this kind of grating consists of a row of connected reflective stripes alternating at two different heights.

#### Reflective phase grating

A phase grating is a bit more complicated to model than an amplitude grating. The main reason for the complexity is the fact that the height in relation to the wavelength determines the irradiance ratio between the different orders. The principle will be explained in two ways. First the same wavefront and phase relations are used as with the amplitude gratings. The second method uses the complete wavefront over the entire surface of the grating.

The example configuration for the explanation, as shown in Figure 7.38, is a special version of a phase grating, where the height difference of the steps in the grating equals exactly a quarter of the wavelength of the perpendicular incoming light. In that case, the orthogonally reflected light coming from a high area will have a  $180^\circ$  phase difference to the reflected light from a low area, as the latter has to travel two times the height of the grating. This optical path difference results in a phase difference of  $\lambda/2$  between the light reflected by both areas. When these areas have an equal surface, the resulting light at the far field in the 0<sup>th</sup> order direction will be cancelled out



**Figure 7.38:** Diffraction at a reflective phase grating, illuminated perpendicular to the surface. Light, reflecting straight back from the bottom of the grating, has to pass two times  $\lambda/4$  more distance than light reflecting from the top. These two will then have  $\lambda/2 = 180^\circ$  phase relationship, so destructive interference occurs at the far field and no light will be reflected in the  $0^{\text{th}}$  order direction. Under an angle the path differences cancel out, so constructive interference occurs, as can be observed from the tangential lines of equal phase.

by destructive interference.

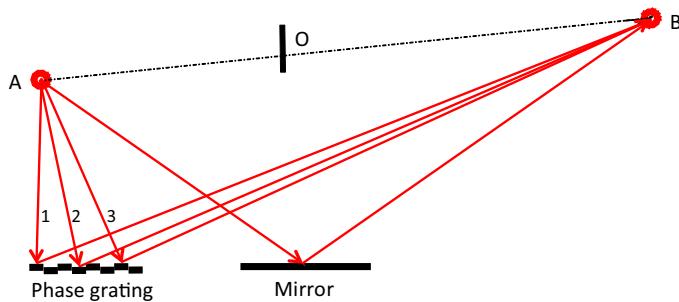
Knowing that energy can not get lost and no dissipation takes place, the light should show an equal constructive interference in another direction and indeed, at a certain angle, the travelled distances will be equal again for all light coming in orthogonal and parting in that diffraction direction as can be seen with the tangential lines of equal phase in the figure.

Identical to the previously explained transmission grating this diffraction angle is equal to Equation (7.28):

$$\vartheta_N = \arcsin \frac{N\lambda}{p} \quad (7.29)$$

This means that also with a phase grating higher order diffraction angles occur, but in practice the geometry of the grating is often chosen such that these higher orders are limited to  $N = \pm 1$  or only have a very low irradiance.

Although only the effect of one point in the middle of each surface is shown, this point is representative for the entire grating as for every point on the low area a corresponding point at the high area can be found that combines in the same way. By integration the total effect can be determined and



**Figure 7.39:** With a reflective phase grating, light can be reflected by a different location than determined by the law of reflection, as long as the phase is equal for the different paths (1,2,3) of the grating.

with the second method to explain the working principle of the transmissive version of a phase grating, the total surface is used.

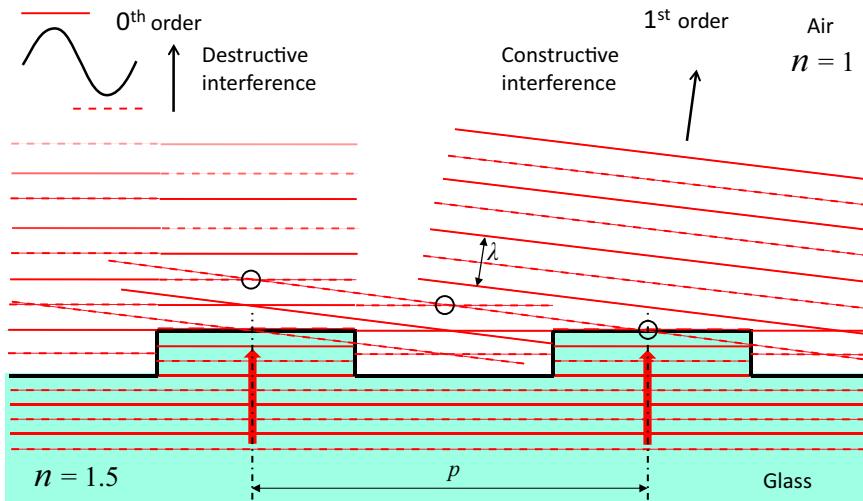
Because of the fact that with a phase grating no absorbing surface is present, the total amount of light is retained.

### Violating the law of reflection?

Several readers might wonder if reflective phase gratings can work at all with a real mirroring surface. With real measurements on these gratings it indeed appears that the reflected light will act different than would be expected by applying the law of reflection that was derived in Section 7.2. Figure 7.39 shows the reflection at a mirror and a grating. At first sight only the middle part of a large mirror is needed for reflection and the photons emitted by the source in other directions are not useful and can never reach the point of observation.

With a grating surface however, this is no longer true as the photons can be reflected in the diffraction orders even with a mirrored surface. This is due to the fact that the simple wave theory is not sufficient to explain everything. Although the method of least time is useful, another method is needed to explain what really happens at a mirrored grating. Again, like also mentioned in Section 7.2, the theory of light acting as particles helps for this explanation.

In a nutshell this theory states that photons in principle can follow all kind of different paths to a target point of observation. The probability of a certain path to this location is however determined by the question whether other photons from the same source, that arrive at that location via different



**Figure 7.40:** Diffraction at a transmission phase grating, illuminated from below perpendicular to the surface. The step size is such that in air ( $n = 1$ ) it equals one wavelength and in the high index material like glass with  $n = 1.5$  it equals 1.5 times the wavelength. In the 0<sup>th</sup> order direction this results in two wavefronts with 180° phase difference that cancel each other out while simultaneously the 1<sup>st</sup> order direction does not show any phase difference.

paths, all have an equal time relation to their temporal periodicity (read, the same phase). For a reflective grating that means that a photon that is reflected at one high step should have the same phase at the point of observation as any other photon from the same source reflected by another high area.

This again is an example of a theory that as such is not understandable, because how can a photon know anything about its colleagues, before it is arrived at the point of observation?

Nevertheless the theory is useful as it perfectly predicts their behaviour.

### Transmission phase grating

A transmissive version of the phase grating can be realised in a transparent plate by a similar surface profile as shown with the reflective phase grating. In this case however the height of the steps is different from the reflective phase grating, in order to obtain the same effect.

When no light in the 1<sup>st</sup> order is allowed, a larger step size is needed in order

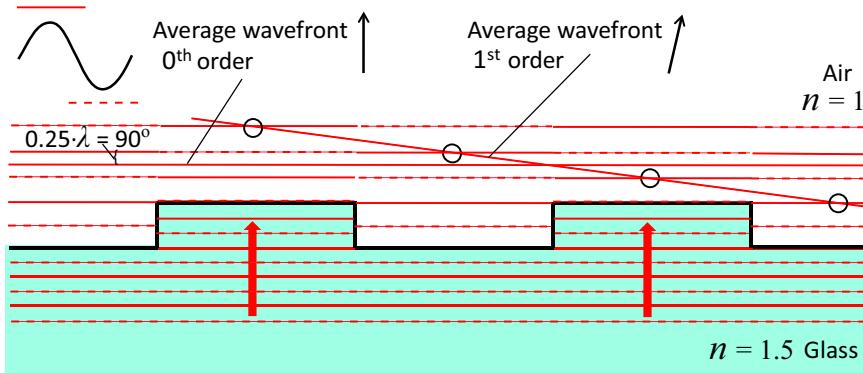
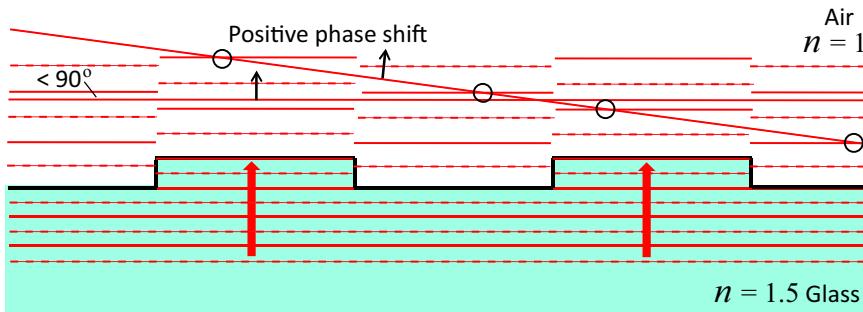
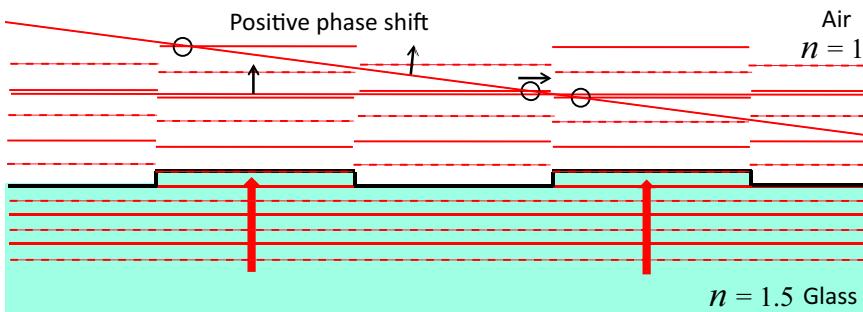
to create the necessary  $\lambda/2$  phase difference between the photons that pass through the area with a shorter optical path length and the photons that pass through the area with a longer optical path length. This means that the height is depending on the refractive index of the transparent material. Figure 7.40 shows such a grating and its principle is explained by using the second method with the full surface. With this method, the transmitted wavefronts are drawn as if they remain flat, instead of a point source with a circular wavefront. This is the extreme other side of the approximation and shows to give the same result. In this example, the step size is chosen such that its height corresponds with one wavelength in air and 1.5 wavelength in the high index material, like glass with  $n = 1.5$ . When observing the 0<sup>th</sup> order the wavefronts from the high and low areas have a 180° phase difference and at the far field these wavefronts will cancel each other out by destructive interference. This is fully comparable with the effect of the reflective phase grating.

### Phase shifting

The transmission phase grating is more easy to visualise in a drawing than the reflective phase grating, because in transmission the incoming light and the diffracted light are at different sides of the grating. For that reason the important phase behaviour of a phase grating at different step sizes is better explained with transparent phase gratings, while the following reasoning is also valid for reflective phase gratings.

It was shown in Equation (7.24) that the phase of two interfering waves with equal magnitude will equal the average of the phase of the two waves. For the phase grating of Figure 7.40 this means that the phase of the resulting wave of the 0<sup>th</sup> order would be 90° different from each interfering wave, even though the amplitude would be zero. The phase of the 1<sup>st</sup> order wave would be in phase with both interfering waves.

It is even more interesting to see what happens, when the step size is different from the above example. This is shown in Figure 7.41. The upper graph shows the previous situation where the 0<sup>th</sup> order is cancelled with 90° phase difference and the 1<sup>st</sup> order is in phase with both waves.

a: Stepsize  $1 \cdot \lambda$  in air,  $1.5 \cdot \lambda$  in glass.b: Stepsize  $0.66 \cdot \lambda$  in air,  $1 \cdot \lambda$  in glass.c: Stepsize  $0.33 \cdot \lambda$  in air,  $0.5 \cdot \lambda$  in glass.

**Figure 7.41:** By varying the step size of the phase grating, the magnitude and phase of the 0<sup>th</sup> and 1<sup>st</sup> order is also varied. A reduced step height (b: and c:) results in a phase advancement of both orders. The irradiance of the orders shifts from 100 % in the 1<sup>st</sup> order at a: to ultimately 100 % in the 0<sup>th</sup> order, when the step size would be zero.

When the step size is changed, as shown in Figure 7.41 b: and c:, the phase of the wave through the high area is less delayed in respect to the wave through the low area. As a result **both orders get a partial positive phase shift**. With a smaller step size, the phase difference between the interfering waves in the 0<sup>th</sup> order direction will be less than 180°. Consequently the magnitude of the irradiance of the resulting wave will no longer be zero, while the phase delay will be less than 90°. At the same time also the waves in the 1<sup>st</sup> order direction will be less perfectly in phase, resulting in a phase difference of the resulting wave and a smaller magnitude of the irradiation. Ultimately when the step size is zero the phase grating will be just a transparent plate again, only transmitting light in the same direction as the incoming light. With the reflective version of the phase grating this ultimate situation would simply be a flat mirror.

As a result of this reasoning, the phase of the 0<sup>th</sup> order will always remain delayed relative to the phase of the 1<sup>st</sup> order diffracted wave. While the delay of 90° from the 0<sup>th</sup> order relative to the incoming light gradually reduces to zero, the phase of the 1<sup>st</sup> order will gradually shift to a positive 90° difference with the incoming light. Regarding the magnitude, following the law of conservation of energy, the sum of the irradiance of the different orders always remains equal to the irradiance of the incoming light as the size of the beams is not affected by the grating. By choosing the right height it is for instance possible to create a grating with the same irradiance for the 0<sup>th</sup> and the two 1<sup>st</sup> order beams. In that situation the phase shift of both orders should be equal. A small approximating calculation shows the effect using Equation (7.26) in a simplified form:

$$I_{r,o} = I_{r,i} \left( \frac{1 + \cos \varphi}{2} \right) \quad (7.30)$$

with  $I_{r,o}$  equaling the irradiance of an output order and  $I_{r,i}$  being the irradiance of the incoming light. When the light going into higher orders is estimated to be about 25 % of the total irradiance, which is a realistic figure, then 25 % is available for each of the three main orders. This level requires the cosine term to be equal to 0.25 for all three orders, resulting in a phase angle of:

$$\varphi = \arccos(2 \cdot 0.25 - 1) \approx 120^\circ \quad (7.31)$$

This means for the 0<sup>th</sup> order that the phase difference between the interfering waves equals 120° and as a consequence the (negative!) phase shift of the 0<sup>th</sup> order after interference becomes half of that value equaling -60° relative to phase of the incoming light. For the 1<sup>st</sup> order the phase is shifted in the

positive direction with the same value of  $60^\circ$  relative to  $0^\circ$ .

The phase difference of  $120^\circ$  in this special phase grating proves to be very useful in the advanced interferometry encoders that will be presented in Chapter 8.

#### 7.4.3.3 Direction of the incoming light

The shown examples all had a strictly defined direction of the incoming beam of light, orthogonal to the surface and one may wonder what happens, when the light approaches the grating from another direction.

In principle, with a reflective grating, a beam with an angle of incidence  $\vartheta$  results in a  $0^{\text{th}}$  order directed with a same angle  $\vartheta$  opposite to the normal fully according to the laws of reflection. In fact as a first-order approximation the geometric effect just linearly combines with the diffraction effect. In reality a large angle of incidence of the incoming light will cancel out certain diffraction orders that would otherwise be directed into the reflecting material while other orders appear at the opposite side of the normal. The interested reader can derive this when drawing the different graphs for other light directions.

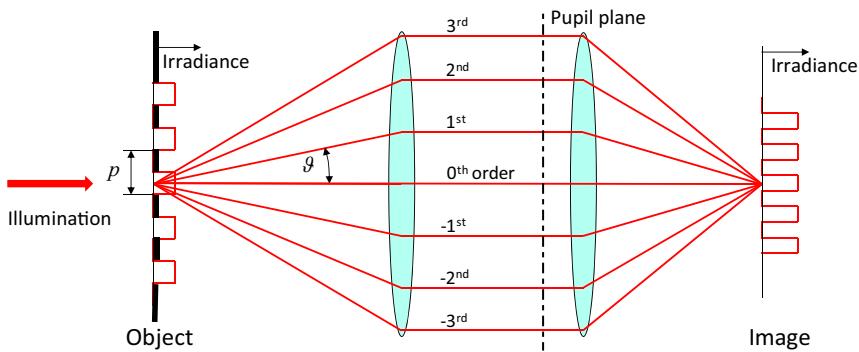
One consequence of this reasoning is that the effect works in two directions. With for instance a reflective amplitude grating, light that enters under an angle corresponding with the  $1^{\text{st}}$  order will be reflected mainly in the direction of the  $0^{\text{th}}$  order. This reverse effect can also be explained by reasoning that the optical path lengths for photons going in the opposite direction are equal, resulting in equal travelling time and an equal phase for their wave property.

With transmission versions of the different gratings in principle the same reasoning is valid al long as the transparent part is planar, without optical strength or wedge form. Otherwise the angle should be corrected for these effects.

#### 7.4.4 Imaging quality based on diffraction

In a purely physical manner, an optical object can be approximated as an accumulation of spatially separated area's with a different irradiance. This irradiance can either originate from local light sources like is the case with a TV-screen or by transmitted or reflected light from a separate source.

It was shown in the previous section on physical optics that separate illuminated lines at equal distance with dark lines in between behave like a

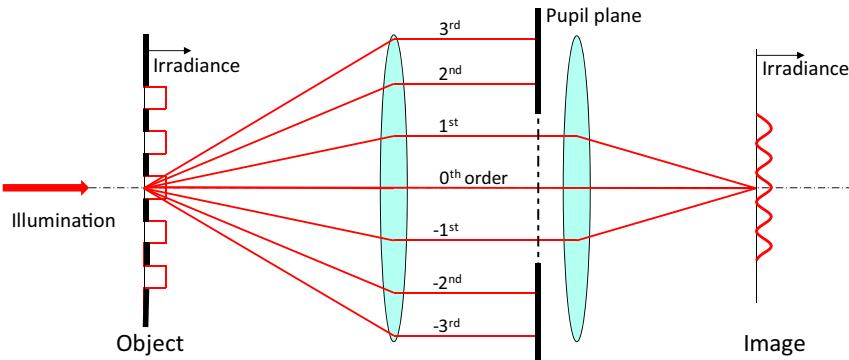


**Figure 7.42:** Imaging of a transparent grating with a square wave irradiance profile illuminated perpendicular to the grating from the left. It shows that the higher diffraction orders that represent the spatial high frequency content of the image are located at the pupil plane further away from the optical axis.

grating and create separate beams of light under an angle depending on the wavelength of the light and the distance of the lines. A small distance between the lines is equivalent to a high density of the lines. This density is expressed in the *spatial frequency*  $f$  of the lines, defined as a number of lines per unit of length ( $1/m$ ). Similar to electrical signals, the spatial frequency distribution of an optical object can be analysed by means of a Fourier series expansion. Application of the Fourier series expansion to a periodic optical signal with a spatial frequency  $f$  results in its sinusoidal harmonics, each with a spatial frequency  $n \cdot f$ . Figure 7.42 shows what this means, when imaging an object that consists of opaque and transparent stripes at equal distance. It is a transparent amplitude grating that is illuminated from the left perpendicular to the grating.

For this example a double-telecentric imaging system is taken for reason of simplicity but the following reasoning is applicable for all imaging systems. Starting at the left side, diffraction orders are generated at the grating that are captured by the optical system. In the pupil plane these orders all cause parallel beams of light, the higher order beam further away from the optical axis than the lower order beam. After the second lens these beams recombine to the image. This image has the same shape as the object at a different size, because of the demagnification in this example. The higher irradiance at the image is caused by the law of conservation of energy, smaller details will show a higher irradiance.

With diffraction theory it was shown that the angle  $\vartheta$  between the diffraction orders equals  $\arcsin n\lambda/d$  which means that a higher spatial frequency

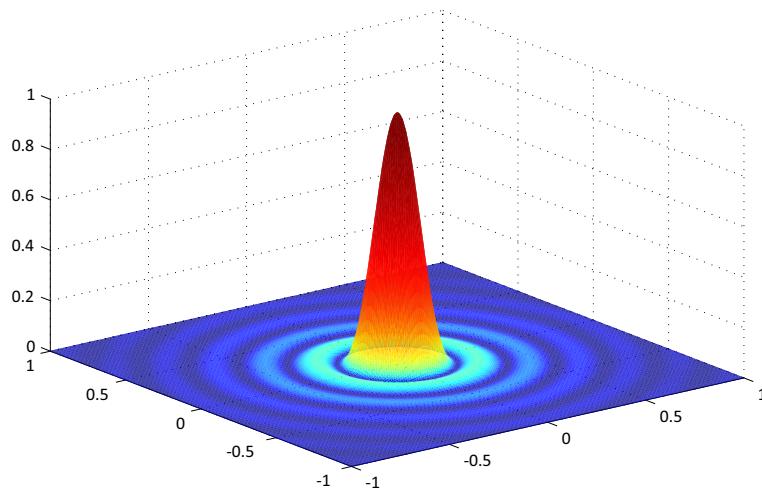


**Figure 7.43:** Imaging of a transparent grating with a square wave irradiance profile with a small aperture stop at the pupil plane showing the spatial low pass filtering of an optical system when blocking the higher diffraction orders.

(smaller  $d$ ) should correspond with a larger angle. This is exactly what is observed when the image side is compared with the object side of the system.

What is not shown, because the figure is erroneous to that respect, is the effect of the missing higher orders that are not captured by the lens. With a real optical system a decrease in steepness and amplitude of the irradiance profile of the image would be observed. This effect is illustrated in Figure 7.43 where a smaller aperture is inserted at the pupil plane in the optical system, cutting off all but the 0<sup>th</sup> and 1<sup>st</sup> order. The diffraction orders created by the square wave irradiance profile each represent a different harmonic of the spatial frequency. The 0<sup>th</sup> order represents the average light irradiance so the DC value of the image,  $a_0$  from the Fourier series expansion. The 1<sup>st</sup> order represents the sinusoidal spatial frequency of the image and the higher orders represent the higher frequency components. Because optical signals are a bit different from electrical signals it is useful to highlight some differences. First of all, negative signal values do not exist for the irradiance as being the magnitude of the electric field squared. This means that when there is an image, there is always an average irradiance value and consequently the 0<sup>th</sup> order is never zero. Secondly the higher optical orders correspond with the odd harmonics from the Fourier series expansion as a square wave consists only of the odd harmonics. In Chapter 2 it was shown that this is true for all signals with a certain symmetry.

The shown examples had a regular square wave pattern but in reality objects never consist of only one spatial frequency. From Fourier analysis it is also



**Figure 7.44:** Airy-disk irradiance profile of the image of an ideal point-source created by a diffraction limited optical system. The fringes around the centre peak are caused by the sharp filtering of the optical system which is equivalent to a higher order dynamic system.

known that a random signal can be approximated by an infinite amount of sinusoidal frequency components.

This leads to the following important statements for optical objects in the spatial frequency domain:

- Any optical object or image can be represented by an infinite amount of spatial frequencies in three dimensions.
- The high spatial frequencies represent the details of the object and they are “encoded” in the area in the pupil plane that is the most distant from the optical axis.
- High spatial frequencies require a large capture angle of the optical system.

In other words, an optical imaging system acts like a spatial low pass filter with a cut off frequency that is determined by the capture angle. A small capture angle results in a lower optical resolution and less sharp images.

An optical system of which the performance is only determined by this diffraction related limitation is called *diffraction limited*. In a diffraction limited optical system all aberrations are smaller than the smallest details in the image. An example of a diffraction limited image is the airy-disk of

Figure 7.44 which is an image of an ideal point-source. The point source with its infinitely small size and infinitely high radiance is the spatial equivalent of an impulse, having an infinite amount of frequencies. In the airy-disk the higher order response of the spatial low pass filtering by the optical system is visible by the shown oscillations around the irradiance peak in the centre, a spatial “dynamic” phenomenon. As soon as the performance of an optical system is not ideally diffraction limited, the airy-disk of the image of a point source will show both a wider and lower middle irradiance peak. The ratio between the peak height of a real optical system and the peak height of a diffraction limited version of this optical system is called the *Strehl ratio*, named after the German physicist Karl Strehl (1864 – 1940). The Strehl ratio is a measure for the quality of an optical system.

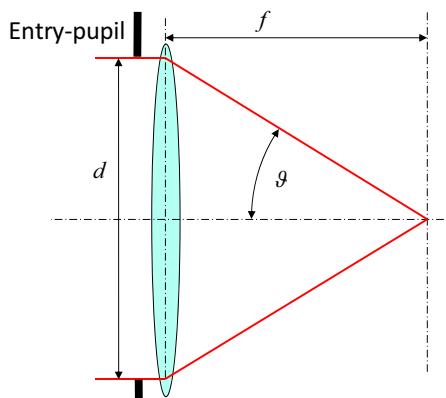
#### 7.4.4.1 Numerical aperture and f-number

Two closely related terms are used in optics to relate the above presented capture angle with the resolution of the imaging system, the *Numerical Aperture* (NA) in professional imaging systems and the *f-number* ( $f/\#$ ) in photography. With the help of Figure 7.45 both terms will be explained. The f-number is the simplest term of the two as it does not take the refractive index into account and equals just the ratio between the diameter of the entry-pupil and the focal length.

$$\langle f/\# \rangle = \frac{f}{d} \quad (7.32)$$

This term is perfectly usable in photographic systems that work in air with objects that are mostly far away and with an image location near the focal plane. The maximum entry-pupil of a photographic lens is almost equal to the diameter of the front lens. Usually the units, also called *stops* differ a factor  $\sqrt{2}$  because the captured light from the object is a function of the surface area of entry-pupil, being squared proportional to its diameter. This means that a stop is a factor two in light irradiance.

Regular photography lenses have an f-number between  $f/2.8$  and  $f/22$ . Often a practical lens design is optimised for aberrations at an f-number around  $f/8$ . This already indicates that these lenses are generally not diffraction limited with the exception of very expensive professional photographic lenses that can be used wide open and sometimes even show f-numbers up to  $f/1$ . It is obvious that these lenses need to be very large to achieve such an opening angle, like those used by sports photographers around football stadiums. It also points convincingly to the inherent limitations in resolution of the small



**Figure 7.45:** Definition of the Numerical Aperture ( $NA = n \sin \vartheta$ ) and the f-number ( $f/\# = f/d$ ) in an optical system. A large NA or a low  $f/\#$  correspond to more captured diffraction orders giving a better resolution. For the illustration the entry-pupil as used for the f-number with photographic lenses is drawn just before the lens. With a telecentric lens the pupil is located at infinity for which reason the capture angle is taken as reference for the NA.

lenses in pocket-size cameras and cell phones. Contrary to the photographic application, professional imaging systems like wafer scanners, that are used for defining the structures in lithography for IC manufacturing, often have both the object and the image close by and show significantly higher capture angles. In this case the Numerical Aperture (NA) is a more suitable quality value and it is defined as follows:

$$NA = n \sin \vartheta \quad (7.33)$$

This term includes the refractive index  $n$  which is related to the fact that the wavelength of light with a specific frequency is shorter at a higher refractive index. As a consequence the angle between the diffraction orders is smaller. In other words an optical system can capture more orders in a high refractive index medium like water or oil than in air.

This is the reason for oil-immersion in microscopes and water-immersion in wafer scanners<sup>6</sup>

The use of the sine instead of the tangent as with the f-number is more practical as it clearly limits the value in air to one as the unattainable

---

<sup>6</sup>The application of immersion fluids in an existing air-based optical design does not automatically increase the NA but only enables the designer to achieve a higher NA with an adapted design that could otherwise not be realised in air.

goal. With an NA in air that is equal to one, the object or image should be positioned inside the lens element which is not possible. With small angles and an refractive index of  $n = 1$  the relation between f-number and the NA can be approximated as follows:

$$\text{NA} = n \sin \vartheta = n \sin \arctan \frac{d}{2f} \approx \frac{1}{2\langle f/\# \rangle} \quad \text{or} \quad \langle f/\# \rangle \approx \frac{1}{2\text{NA}} \quad (7.34)$$

The numerical aperture determines the maximum resolution of a lens, also called *Critical Dimension* (CD)<sup>7</sup>, with the following relation:

$$\text{CD} = k_1 \frac{\lambda}{\text{NA}} \quad (7.35)$$

with  $\lambda$  the wavelength of the light in vacuum because the refractive index has already been taken into account in the NA value and  $\lambda$  is mostly known from the light source. The term  $k_1$  is named just like that, the *k-one factor* and equals around 0.5 for a not optimised system.

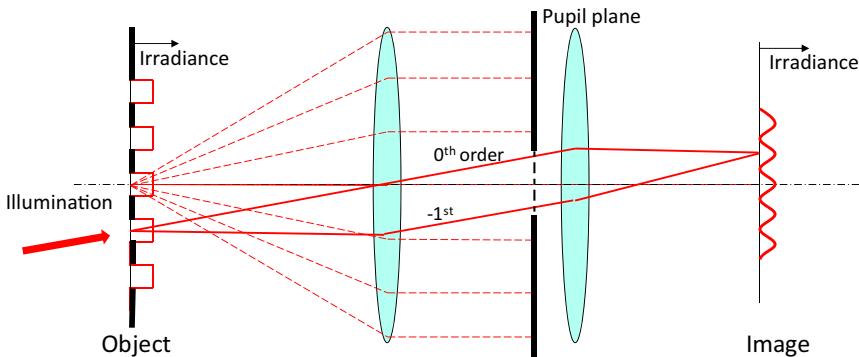
With special “tricks” that mainly consist of illumination methods at the object side, like shown in Figure 7.46, the system is able to image at a higher resolution. The figure shows as an example the effect of illumination from a direction under an angle with the optical axis, causing a tilting of all the diffraction orders. With a perpendicular illumination the higher orders would be blocked at the aperture stop (the dashed lines) but with this angular illumination both the 0<sup>th</sup> and one of the higher orders is able to pass the aperture, giving enough optical information to create an image.

When the object would be illuminated from two directions under the same angle with the optical axis or even from all directions in three dimensions (a cone) under the same angle, called *annular illumination* even both 1<sup>st</sup> orders would be captured. Further also contrast enhancement methods at the image side help to lower this value. The real minimum value for  $k_1$  is 0.25 as this is the situation where not even a fraction of the 1<sup>st</sup> order is captured any more with whatever illumination method.

The mentioned methods for achieving a low  $k_1$  are applied in waferscanners as these are not aimed at reproducing a nice representative image of the object like in photography. Instead, they are designed to produce an image of a certain structure by optimising all related elements. In that case the

---

<sup>7</sup>The term (NA) and (CD) are not written as a variable in one italic symbol and strictly it should be given a different real variable name but the used notation is quite common in those industries where it is relevant. For this reason the NA and the CD are noted with straight capitals.



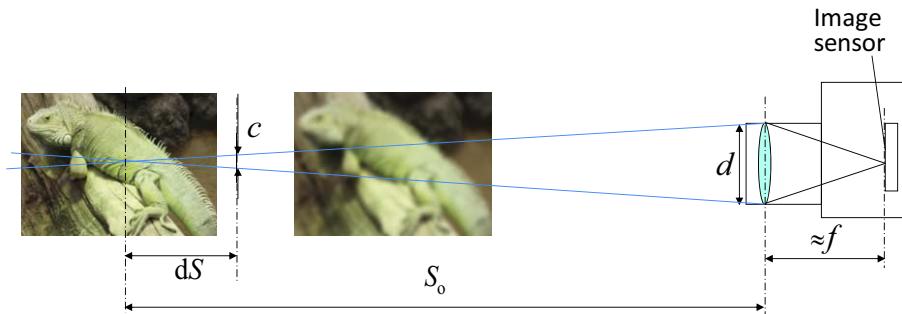
**Figure 7.46:** Illumination under an angle with the optical axis enables imaging in case the aperture is too small for capturing both 1<sup>st</sup> order angles (dashed lines). The changed angles enable the capturing of the 0<sup>th</sup> and one of the 1<sup>st</sup> orders.

object, also called the mask or *reticle*, can be adapted to make optimal use of the possibilities of the entire optical system. The contrast enhancement techniques at the image are realised in the photosensitive resist that sharply distinguishes between different levels of irradiance within only a few percent. It is needless to say that this requires a tight control of the average light level, called *Dose-control*.

#### 7.4.4.2 Depth of focus

The tolerance in positioning of an object or an image sensor along the optical axis in respect to the optical system has many names and *depth of focus* is one of them. Other names like *depth of field* relate often to the tolerance in position of the object while depth of focus is used at the image side. The use of the word “field” can be confusing as it also relates to a force field and in imaging the object and image plane are also often named a field. Furthermore both positions have a simple relation depending on the focal length of the optical system. For these reasons the name depth of focus (DOF) is used in this book.

When the position of the sensor to the lens is fixated, the object is only imaged sharply around a small area that depends on several parameters. This tolerance depends primarily on the aperture angle. This means that a high NA or low f-number, as would be necessary to achieve the maximum resolution possible, will result in a low depth of focus. With a large aperture angle only a little displacement away from the focal point will rapidly result



**Figure 7.47:** The depth of focus ( $dS$ ) of a photographic camera depends on different parameters: the distance of the subject to the lens ( $S_o$ ), the minimal resolution ( $c$ ) and the f-number ( $\langle f/\# \rangle = f/d$ ).

in a larger luminous spot proportional to this displacement. In non critical optical systems like in photography where the minimum resolution ( $c$ ) is higher than the diffraction limited value, the allowable error ( $dS$ ) in the focal displacement of the object ( $S_0$ ) in relation to the diameter ( $d$ ) of the entry-pupil can be derived from simple geometry as shown in Figure 7.47:

$$\frac{dS}{S_o} = \pm \frac{c}{d} \quad (7.36)$$

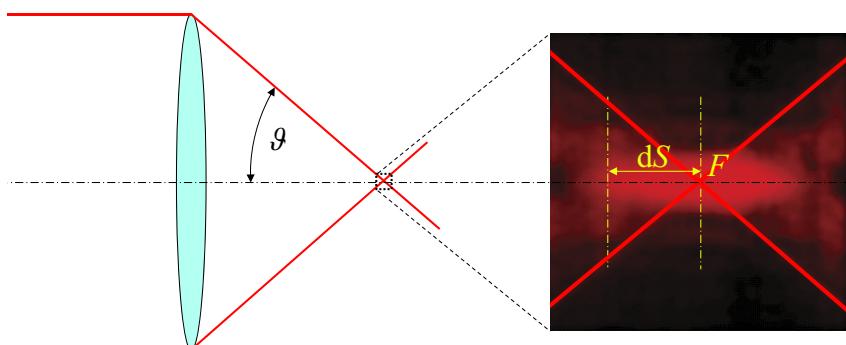
With the f-number ( $\langle f/\# \rangle = f/d$ ) the following relation for the maximum allowable focal error is derived:

$$dS = \pm \frac{c S_o \langle f/\# \rangle}{f} \quad (7.37)$$

A large distance, a high f-number, a large minimum resolution value and a small focal distance all result in a high depth of focus. This corresponds with the experience of photographers that a telephoto-lens shows a smaller depth of focus than a wide-angle lens that a high f-number is required for landscape pictures with objects both nearby and far off that have to be imaged sharp. It is also the reason that the integrated cameras in inexpensive cell-phones have often a wide angle lens because of they fail a focusing motor.

The situation at the image side of an optical system is shown in Figure 7.48. Although the reasoning is identical as for the object side, the distance from the image to the lens is in most cases much smaller so also the allowable error in the positioning of the image sensor of a photographic camera is far smaller. Fortunately this sensor is flat and with automatic focusing it is possible to correct for misalignment.

More interesting is the shown effect at the focal point when observing a



**Figure 7.48:** Depth of focus with diffraction limited optics. The enlarged picture at the right shows the irradiance at the focal point. The profile of an airy-disk can be observed at the cross section of the focal point, with dark and light areas next to the central spot. Further it also shows that the size of the spot remains about the same over a larger range along the optical axis than expected from the crossing lines, resulting in a depth of focus ( $dS$ ) of about 2 – 3 times the spot size.

diffraction limited image. In that case it is not possible to simply derive the tolerance from geometry of the two crossing most outer rays as around the focal point the wave character shows a very interesting effect.

It was shown in the previous section that the size of the image at the focal point can never have an infinitely small size because of the size of the airy-disk that is determined by the numerical aperture.

In spite of this minimum size of the spot in the centre, the maximum irradiance area spreads over a larger area in the direction of the optical axis than would be expected from purely geometrical analysis. This results in a larger focal range according to the following relation:

$$dS = \pm \frac{n\lambda}{NA^2} \quad (7.38)$$

In Chapter 9 it will be shown that because of this phenomenon an optical resolution in the order of 50 nm can be achieved with a depth of focus in the order of 100 nm, when using immersion optics with a very high numerical aperture of  $\approx 1,35$ .

## 7.5 Adaptive optics

In the past decades, the principles of active control are increasingly applied in complex imaging systems to achieve the ultimate performance. This mechatronic field of *adaptive optics* has proven its value in astronomy and high precision optics for IC lithography but also in medical instrumentation it became a method to achieve results that were hitherto deemed impossible.

In this section one of the main causes of a less than optimal optical performance of an imaging system is presented which is the generally strong temperature dependency of the optical path of the light.

It will be shown that the resulting problems can be reduced by inserting active optical elements such as deformable mirrors with a suitable measurement and control system.

### 7.5.1 Thermal effects in optical imaging systems

Optical systems operate by virtue of differences in the velocity of light, represented by the refractive index of the transparent material. Unfortunately the refractive index is largely influenced by temperature. The impact on the refractive index by temperature is mainly caused by the changing density of most materials as function of temperature. In most cases an increasing temperature will cause a decreasing density. With a lower density there are fewer atoms to slow down the light over the trajectory which is equivalent to a lower refractive index.

This however is not always the case. Especially in glass, the refractive index can increase or decrease as function of rising temperature due to the composition with special elements. Next to this direct effect on the refractive index, the size of an object depends on the temperature. For a lens this means that also its curvature is not constant.

To examine these effects in a qualitative manner the previously introduced lens-maker's equation can be used:

$$\frac{1}{f} = (n - 1) \left( \frac{1}{R_1} - \frac{1}{R_2} \right) \quad (7.39)$$

At first sight it might seem possible to select a glass type with such a positive  $dn/dT$  that it compensates the effect of the increasing radius by the larger lens size at a higher temperature.

In reality however often this choice is mostly not available while the effect on geometry and refractive index of otherwise preferred materials can be



**Figure 7.49:** The car looks as if standing in water while being in the middle of the Mojave desert at 40 °C. This mirage is caused by a lens effect that refracts light from the sky towards the observer. This lens effect is created by the hot air with a low density close to the ground. Also the turbulence caused by the thermally induced motion of the air is clearly visible at the telephone poles while the image of the approaching car at the horizon is even hardly distinguishable.

quite different.

With for instance the professional lenses for IC lithographic exposure equipment working at the Deep UV wavelength of 193 nm, it is necessary to use Fused Silica as lens material that shows an large temperature dependency of the refractive index  $dn/dT = 15 \cdot 10^{-6}$  while the thermal expansion coefficient is only  $\approx 0.5 \cdot 10^{-6}$ .

This means that one °K of temperature could shift the focal plane of a lens with a focal length of 10 mm with 150 nm which is not acceptable for such a precision optical system. Even worse is the situation where the temperature is not constant over the optical system due to absorption of light causing additional aberrations.

Another well-known example of thermally induced optical errors is the effect of temperature differences in air like shown in Figure 7.11 in Section 7.2.2. It is the cause for the famous phenomenon of a *mirage* as shown in Figure 7.49.

### 7.5.2 Correcting the wavefront

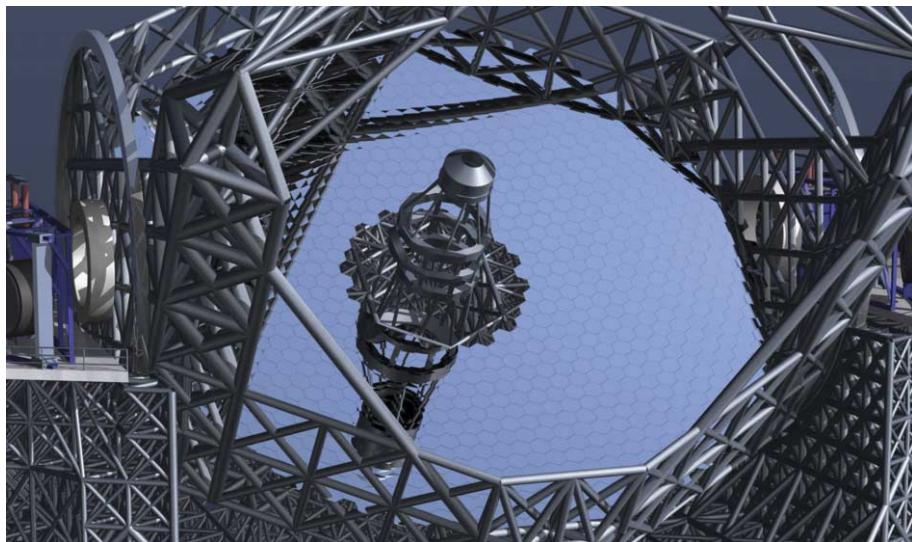
The pupil plane inside an optical system is shown to contain all the spatial information about the object that is imaged. This information is stored in the different diffraction orders that all propagate parallel for any point on the subject.

The wavefront at the pupil plane should be flat because of these parallel rays. This means that at the pupil plane aberrations can be distinguished as deviations from this flat wavefront. It was further also shown that the information on the details is spread over the entire pupil plane with the finer structures more distant from the optical axis. Adaptive Optics is based on the principle to introduce a controllable optical element in the optical path that corrects these deviations from the ideal wavefront after measurement by a wavefront sensor.

The first application of adaptive optics has been in terrestrial telescopes where the turbulence and temperature effects in the atmosphere determine changes in the refractive index leading to wavefront errors. Especially with the ever larger size of these telescopes, necessary to capture more light and increase the resolution, the influence of the air becomes dominant. As an example Figure 7.50 shows the E-ELT, the Extreme Large Telescope of the European Southern Observatory that is planned to be finished around 2018. The largest mirror has a diameter of  $\approx 40$  metres. It is impossible to manufacture such large mirrors in one piece and it became customary for telescopes to compose such a large mirror from a multitude of smaller elements. The main mirror of the E-ELT consists of around 1000 smaller mirror elements of 1.4 metres wide. Each of these mirror-elements is precisely positioned relative to the other mirrors, such that the light from each element is in phase with the light of the other elements. Only under that condition the combined mirrors will act like one large mirror with one consistent wavefront.

The precision of this matching needs to be better than  $\lambda/20$  which is about 15 nm, because light with wavelengths as short as 300 nm has to be imaged. It is well understandable that a wavefront that passes more than 50 km of atmosphere can easily become distorted by more than this 15 nm and for that reason two of the four other mirrors of the telescope have an actively controlled surface shape.

In the following two sections, first the Zernike modes are introduced as a mathematically description of the wavefront aberrations, while in the second section the principle of operation of adaptive optics is further explained.



**Figure 7.50:** The Extremely Large Telescope of the European Southern Observatory (E-ELT) has a main mirror with a diameter of just less than 40 metres that consists of  $\approx 1000$  segments with a diameter of  $\approx 1.4$  metres that all have to be aligned with an accuracy of  $\lambda/20 \approx 15 - 50$  nm. Two adaptive secondary mirrors are used to correct for wavefront errors by atmospheric disturbance. (courtesy of ESO)

### 7.5.2.1 Zernike modes

A deviation from an ideal wavefront can mathematically be described in orthogonal polynomials, called *Zernike polynomials*, that are defined in a polar coordinate system over a circular plane normalised to radius 1 (unit disk) around the optical axis. The full waveform deviation is then approximated by the summation of a plurality of singular shapes (modes) of a different order and magnitude. This approximation is comparable with the Fourier analysis of signal waveforms and modal shapes of a dynamic system. The different Zernike polynomials or Zernike modes that define these singular shapes are named after the Dutch scientist and Nobel prize winner Fritz Zernike (1888-1966), who defined them. Next to wavefronts, also the surface of optical elements and the image shape in the field plane could be described by these polynomials, but mostly the use of Zernike modes is limited to the pupil plane where the ideal wavefront should be flat. There are even and odd Zernike polynomials, where the even ones are defined

as:

$$Z_p^q(\rho, \varphi) = R_p^q(\rho) \cos(q\varphi) \quad (7.40)$$

and the uneven ones as:

$$Z_p^{-q}(\rho, \varphi) = R_p^q(\rho) \sin(q\varphi) \quad (7.41)$$

where both  $p$  and  $q$  are non negative integer numbers,  $p - q$  are even and  $p \geq q$ .

The angle  $\varphi$  is the azimuthal angle, between a reference vector and another vector that originates in the centre and points towards the location of interest. The value  $\rho$  is the normalised radial distance on the unit disk with diameter of 1.

The radial polynomials  $R_p^q(\rho)$  are given by:

$$R_p^q(\rho) = \sum_{k=0}^{(p-q)/2} \frac{(-1)^k (p-k)!}{k!((p+q)/2-k)!((p-q)/2-k)!} \rho^{p-2k} \quad (7.42)$$

This equation results in the following list of radial polynomials for the first 21 Zernike modes, where  $p + q$  is the order of the Zernike mode<sup>8</sup>:

$$\begin{aligned} R_0^0(\rho) &= 1 \\ R_1^1(\rho) &= r \\ R_2^0(\rho) &= 2\rho^2 - 1 \\ R_2^2(\rho) &= \rho^2 \\ R_3^1(\rho) &= 3\rho^3 - 2\rho \\ R_3^3(\rho) &= \rho^3 \\ R_4^0(\rho) &= 6\rho^4 - 6\rho^2 + 1 \\ R_4^2(\rho) &= 4\rho^4 - 3\rho^2 \\ R_4^4(\rho) &= \rho^4 \\ R_5^1(\rho) &= 10\rho^5 - 12\rho^3 + 3\rho \\ R_5^3(\rho) &= 5\rho^5 - 4\rho^3 \\ R_5^5(\rho) &= \rho^5 \end{aligned} \quad (7.43)$$

<sup>8</sup>Several other definitions of these polynomials exist that differ in details, for instance by normalising on maximal values at the edge of the circle, like 1 as used here. Also the units can relate to metric values like [nm] or wavelengths. Further the order is often mentioned as "p" equalling the *radial order* and "q" equalling the *azimuthal order* and the sequence of numbering is either simply based on an increasing radial order or on a more complicated expansion with spherical terms for all quadratic numbers. All these definitions are correct as long as maintained consequently in all communication.

Is important to iterate another part of the definitions of a wavefront for the explanation of the meaning of these modes in a descriptive form. A wavefront is a representation of the light coming from an infinitely small spot while the photons over the wavefront have a fixed phase relation or equal timing when starting at the same moment. Delaying part of the photons will distort the wavefront and as a consequence the direction of the photons is changed. This all means that each point on the object creates a wavefront that is flat at the pupil plane and can be distorted by errors in the imaging system.

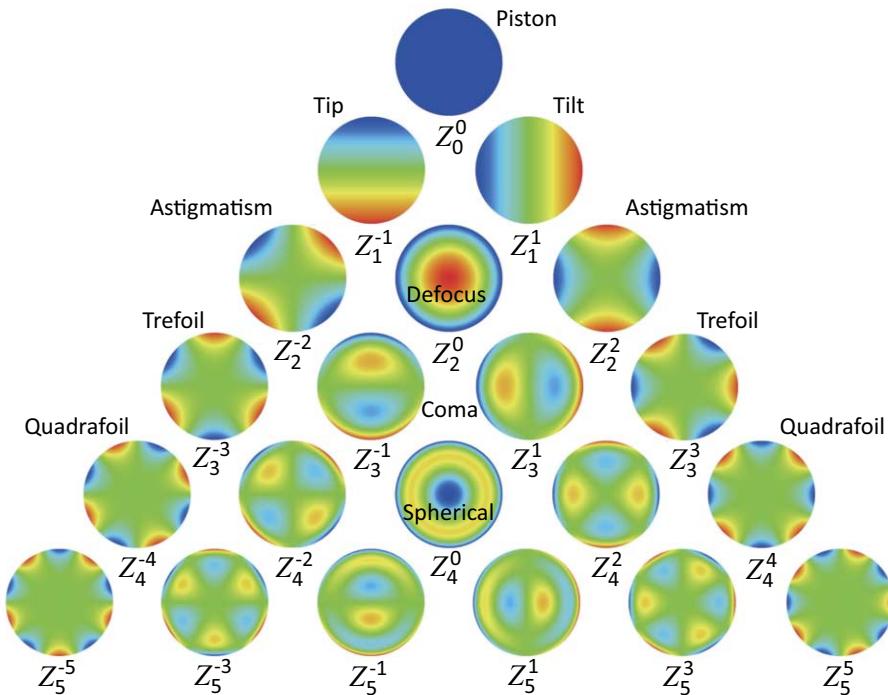
In Figure 7.51 a graphical overview of the first 21 Zernike modes is given to visualise the different shapes. The upper image corresponding with  $Z_0^0 = 1$  is the plane wave term, equal to the mean value of the wavefront. It represents a constant delay of all photons in the wavefront, caused by for instance a flat piece of glass. One step down, where  $p = 1$ , the Zernikes represent the effect of tilting around the vertical or horizontal axis. These effects can be caused by a wedge shape piece of glass, called a *prism* and means that the main direction of the wavefront is changed over a tilting angle. The third row with  $p = 2$  shows in the middle the effect of defocus which causes a parabolic wavefront in the pupil plane and is represented by  $Z_2^0$ . As a consequence of the defocussing, the inner rays will be either delayed in respect to the outer rays or vice versa and they will no longer coincide on the same focal point on the image. The image becomes unsharp.

In the section on geometric optics three aberrations were introduced, spherical aberration, astigmatism and coma. The examples were shown with positive single lens elements where the pupil is determined by the size of the positive lens itself. For linking the aberrations with the images in the figure one can imagine the pupil plane with parallel rays to be somewhere inside the lens element. The primary versions of the three mentioned aberrations are represented by the 4<sup>th</sup> order Zernike terms.

Spherical aberration is represented by  $Z_4^0$  and a cross section through the middle looks like a double sine wave. When comparing with Figure 7.19 the outer and inner rays are delayed in respect to the rays in between. It was shown in that figure that it should be corrected by adding more glass, meaning more delay, in that intermediate area.

Astigmatism is represented by  $Z_2^{-2}$  and  $Z_2^2$  and can be seen as a saddle shape. When comparing with Figure 7.21 it is shown that rays in one plane are delayed in respect to rays on the optical axis while rays in an orthogonal plane are in advance to the rays at the optical axis.

Coma is represented by  $Z_3^{-1}$  and  $Z_3^1$  and its cross section through the low and

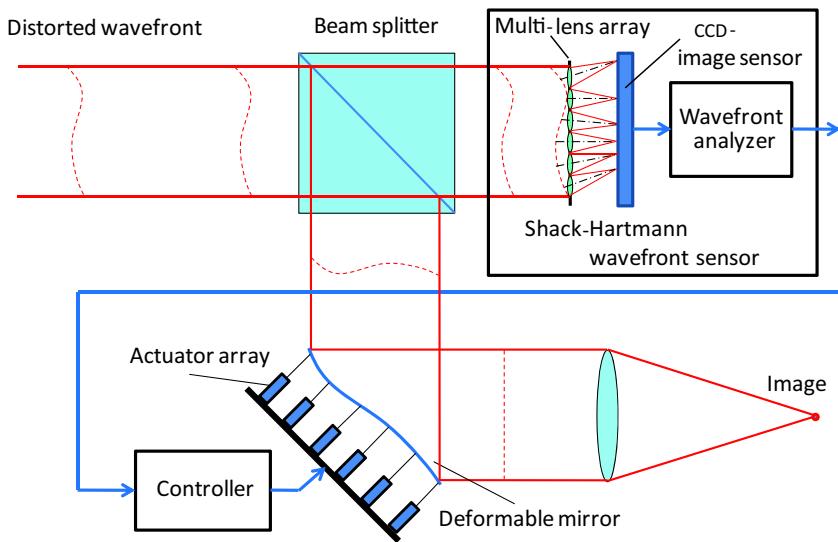


**Figure 7.51:** Height map of the first 21 Zernike modes describing deviations of an ideal flat surface in a systematic manner. Generally deviations of wavefronts can be modelled as a combination of several Zernike modes with different magnitude. In this chart blue corresponds with 1, green with zero and red with -1.  
(courtesy of Claudio Rocchini)

high area looks like a single sine wave. When comparing with Figure 7.22 this is less easy to imagine and with the higher order terms it is not useful to even try, but one thing is certain, the higher the Zernike terms the more refined the correction mechanism should be.

Several of the higher order terms can be seen as high spatial frequency versions of the described primary aberrations. For instance  $Z_4^2$  and  $Z_4^{-2}$  represent secondary astigmatism and  $Z_5^1$  and  $Z_5^{-1}$  is secondary coma.

The other Zernike modes also have their own name based on their shape like *trefoil* for  $Z_3^3$  and  $Z_3^{-3}$ , *quadrafoil* for  $Z_4^4$  and  $Z_4^{-4}$ , *pentafoil* for  $Z_5^5$  and  $Z_5^{-5}$  and consequently  $Z_5^3$  and  $Z_5^{-3}$  are called *secondary trefoil*.



**Figure 7.52:** Principle of adaptive optics with feedforward compensation. By inserting a beam splitter in the pupil area the distorted wavefront can be measured and corrected by a deformable mirror. Notice that the deformation magnitude of the deformable mirror has to be less than the distortion magnitude of the wavefront depending on the angle of incidence.

### 7.5.2.2 Adaptive optics as correction mechanism

Based on these mathematically described deviations of an optical system, adaptive optics is developed as a mechatronic method where the wavefront errors in the pupil plane are measured and corrected by means of an actively controlled optical element and a suitably tuned control system.

In principle one could consider automatic focusing of a photo camera as a kind of adaptive optics and indeed it shows all elements of an active controlled system including sensor, actuator and controller. Generally however the term adaptive optics is reserved for those applications where the curvature of optical elements themselves is actively controlled in order to reduce the wavefront errors.

### 7.5.3 Principle of operation

The principle of operation of adaptive optics is shown in Figure 7.52 which shows a feed forward compensation method. The light with the distorted wavefront enters the system at the upper left side, as shown by the dashed line in the pupil area where the light is running parallel.

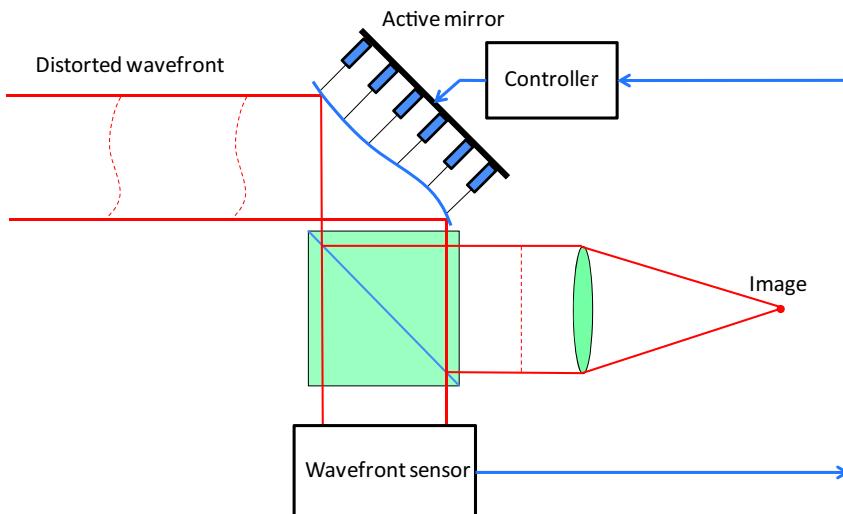
As a first step a *beam splitter* splits the light in two paths. The irradiance ratio between these paths is not important for the working principle but one can imagine that with an expensive telescope it would be a waste to send too many photons to the wavefront sensor.

The wavefront is analysed with a wavefront sensor that is based on the principle that light propagates in a direction orthogonal to the wavefront. The most simple version of a wavefront sensor consists of a plate with tiny holes where the light through the holes will illuminate a CCD image sensor like used in a digital camera that is positioned behind the plate. The position of the illuminated areas is determined by the direction of the light orthogonal to the local wavefront. With a suitable algorithm the wavefront can be reconstructed from the position of the spots on the CCD sensor. This wavefront sensor is called the *Hartmann sensor* after the German physicist Johannes Franz Hartmann (1865 – 1936) who invented the principle. It was later improved by The American physicist Roland Shack to the *Shack-Hartmann wavefront sensor* by replacing the tiny holes by an array of small lenses that are shown in the figure. In the figure this principle is indicated in an exaggerated way, where the distorted waveform results into the mentioned spatially shifted pattern of spots. Although only five lenslets are drawn, in reality many more lenslets are used depending on the required spatial resolution. This refined *multi-lens array* covers two degrees of freedom in order to precisely analyse the wavefront over the entire pupil plane.

The obtained wavefront information is used in a controller to adapt a deformable mirror in such a way that it compensates the waveform distortion by introducing Zernike terms with an equal magnitude as the distorted wavefront but with an opposite sign.

The deformable mirror is equipped with a multitude of actuators that are capable of covering all relevant Zernike terms, comparable to the multi-lens array of the wavefront sensor. At the end of this section an example of such an actuated mirror will be shown in more detail.

It should be noted that the deformation of this correcting mirror has to be less than the magnitude of the distortion of the wavefront. This is caused by the angle of incidence of the light in respect to the surface of the mirror. When the mirror surface is orthogonal to the incident light the optical path



**Figure 7.53:** Principle of adaptive optics with feedback control. The mirror is controlled such that the error on the wavefront sensor is kept as close as possible to zero.

difference of light reflecting on the mirror is twice the displacement of the surface. When the surface is tilted under an angle of  $45^\circ$  as in the shown examples the timing difference of the light  $\sqrt{2}$  times the displacement. A larger tilt angle will require a larger deformation magnitude.

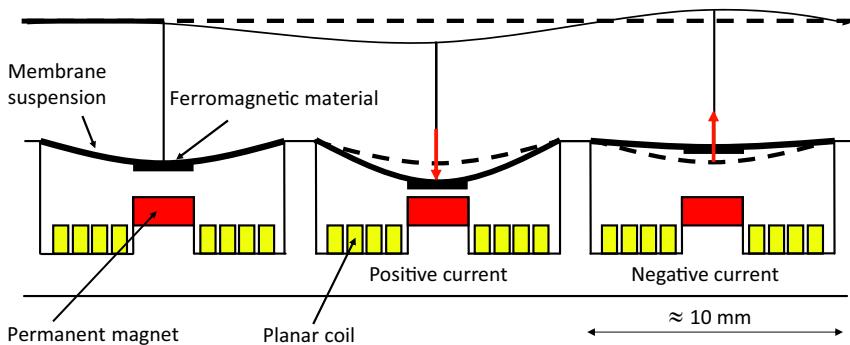
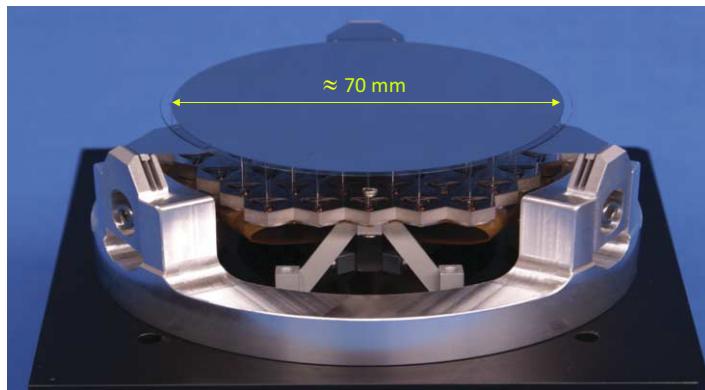
Although this kind of feedforward compensation is used in practice, it requires a precisely predictable behaviour of the actuated mirror and a flawless sensing of the wavefront over a large range. This is only possible with frequent calibration.

For that reason more often a second option is applied by “reshuffling” the parts and creating a real-time feedback loop as is shown in Figure 7.53.

With feedback control the light is reflected first at the active mirror followed by the beam splitter with the sensor and the imaging part.

The mirror is controlled in such a way that the deviations as measured in the sensor are as close as possible to zero. This zeroing tracking-control approach has as benefit that the sensor only needs to be calibrated for a plane wavefront and that the actuators on the mirror only have to be linear enough to not impair the stability of the feedback loop.

In principle all controlled channels from one measured wavefront location to the corresponding actuation part are interconnected and mutually interfere with each other both mechanically and optically so the control algorithm is not obvious. Especially when the errors in large space telescopes need to



**Figure 7.54:** Actively controlled mirror with “Hybrid” permanent magnet biased actuators. The actuator at the left has no current, in the middle the current direction increases the permanent magnet flux and at the right the current direction decreases the flux of the permanent magnet. (courtesy of Roger Hamelinck, TU/e)

be corrected it appears that for a 40 metre telescope around  $10^5$  actuators would be needed. For real-time control this amount of channels requires a multitude of more than  $10^4$  FPGA processors when using classic SISO control. For this reason a real MIMO control system with distributed control algorithms is preferred which takes the actions of the neighbouring elements into account.

Many possibilities have been investigated to realise controllable mirrors, often with piezoelectric or electrostatic actuation of the thin plate that acts as the deformable mirror.

As an example, Figure 7.54 shows a design that was realised at the Eindhoven University of Technology. This configuration uses permanent magnet biased reluctance actuators to determine the shape of the mirror.

This adaptive-optics mirror-system consists of several layers. The upper layer is a thin sheet of glass with a suitable reflection coating that acts as the mirror. It is connected by thin strands of steel wire to the middle layer, the moving part of the actuator, consisting of ferromagnetic pieces of metal that are suspended by a membrane. The lower layer, the stationary part, consists of a base plate with coaxial permanent magnets and flat wound coils that create a magnetic field that can be modulated by the current in the coils. Without current the permanent magnet secures a certain pretension. Depending on the current direction, the attraction force to the ferromagnetic piece of metal is reduced or increased. Due to the coupling with the thin wires, the mirror can adapt its shape according to the average position of the actuator without introducing discontinuities at the connection points while also the heat transfer from the actuators to the mirror is minimised. Only a small part of a larger mirror is shown. In the real system the actuator part will be assembled from a multitude of the shown hexagonal substructures but the mirror itself will be one larger surface.

**Side note:** This version of the “hybrid” actuator has the permanent magnet inserted in series with the flux path of the coil. According to the theory, this design increases the reluctance of the magnetic flux induced by the current, resulting in a reduced energy efficiency.

A more preferred hybrid actuator design with the magnet outside the flux path of the coil would require a symmetric configuration with an additional part on the other side of the membrane suspension as described in Section 5.3.4. Such a design would complicate the layout excessively and also the benefits of a more optimal magnetic configuration are limited because of the small size per actuator with its relatively larger reluctance of the air gaps. For this reason the shown design is more optimal when all factors are taken into account, including manufacturability.

# Chapter 8

## Measurement in mechatronic systems

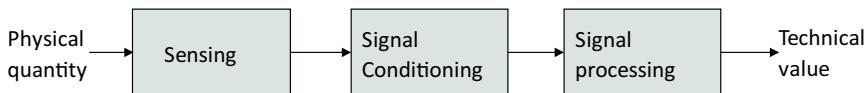
The precision of mechatronic systems can not be better than the measurement accuracy of all relevant parameters that are needed to control and adapt the behaviour of the system. The other elements in the system are never capable of correcting unknown measurement errors.

The field of *metrology* deals with measurement in general and is in itself quite immense, even when only dealing with mechanical quantities. One of the reasons for this wide scope is the economic need for undisputed quantities when trading goods. Metrology has been practised from the very moment that people started to exchange valuable items and this has resulted in agreements and rules on a global scale regarding *traceability*, the possibility to relate a measurement value to an agreed standard like the kilogram and the metre. For the same reason as with the other chapters, this chapter has to remain limited to the most relevant measurement items for precision motion systems, being the measurement of forces, position and motion. This includes sensors for stress, velocity and acceleration, based on physics principles like piezoelectricity, electromagnetism and optics.

This chapter is divided in seven sections.

Section 8.1 is a more general introduction of the basic principles and modelling of measurement systems. Also some definitions from the traceable metrology field will be introduced, especially regarding measurement uncertainty.

Section 8.2 presents the statistics around random errors by introducing an important statistical method, called Dynamic Error Budgeting. This



**Figure 8.1:** Measurement systems can be modelled as a series of three elements.

The sensing part translates the physical signal into an electrical signal.  
The signal conditioning adds robustness to the sensor signal and the  
signal processing translates the signal into a quantitative value.

method enables a mechatronic designer to determine the total error in a dynamic system from different error source contributions. While statistics are very important for the analysis of measurement errors, this method is also applicable for the estimation of positioning errors in a mechatronic motion system.

Section 8.3 concentrates on the electronic signals in the most sensitive elements of a measurement system. Different sources of errors will be distinguished and methods are presented to reduce the negative effects.

Section 8.4 introduces dedicated electronic circuits that are used to add robustness to weak signals. It strongly builds on the basic theory of operational amplifier circuits that was presented in Chapter 6.

Section 8.5 presents the methods to create digital data from the analogue signals in a reliable way.

Section 8.6 is the first of three application sections and concentrates on sensors in short range position measurement systems.

Section 8.7 introduces sensors for the dynamic measurement of mechanical quantities like force, velocity and acceleration.

Section 8.8 presents the important long range optical position measurement systems with encoders and laser interferometry.

## 8.1 Introduction to measurement systems

The function of a measurement system is to translate a physical quantity, the *measurand*, into a representative engineering quantity. Depending on the application, this can be an analogue or a digital representation. For example the pointer of a speedometer in a car is a typical analogue representation, while the numerical reading of a fuel pump in a gas station is an example of a digital representation.

A basic measurement system for controlling a process or function can be modelled to consist of 3 successive functional elements<sup>1</sup> as shown in Figure 8.1 that each have very specific properties.

- A sensing element that converts a variable quantity of a physical phenomenon (physical signal) into an analogue electrical signal.
- A signal conditioning element that adds robustness to the analogue signal by means of amplification such that it can be transferred to another location without risk of disturbances by interfering signals.
- A signal processing element that converts the analogue signal into a different analogue or a digital value that is adapted to the requirements for active control of a larger system and process.

Examples of sensors are a thermocouple that creates a voltage as function of temperature, a light sensitive diode that converts a stream of photons into an electric current and a moving coil in a magnetic field that gives a voltage as function of the velocity. Low-power signal amplifiers are typical examples of signal conditioning elements and an analogue-to-digital converter is a typical signal processing element.

### 8.1.1 Errors in measurement systems, uncertainty

All elements of a measurement system show deviations in their intended performance. These deviations lead to an overall difference between the measured quantity of the measurand and the *real value* that would have been measured with a perfect instrument. This difference, called *measurement error* is the subject of research of many scientists.

<sup>1</sup>In general literature on measuring a 4<sup>th</sup> element is often added that describes the data representation part by means of a display, an analogue pointer, printer or plotter. Because of the limitation to measurements for controlling a dynamic process without an external display, this element is omitted here.

In official *traceable* metrology for economic purposes several terms are defined to classify and deal with measurement errors. These terms are laid down in the "Guide to the Expression of Uncertainty in Measurement" (GUM) and the "International vocabulary of metrology, basic and general concepts and associated terms" (VIM) under responsibility of the "Joint Committee for Guides in Metrology" (JCGM) where the "Bureau International des Poids et Mesures" (BIPM) is participating as international standards organisation. This part of metrology is less relevant for mechatronic systems, because most controlled motion systems rather rely on internal references than on external traceable, often also called *absolute standards*. Nevertheless it is useful to be aware of these conventions when using the different terms in communication with others, for which reason some of these definitions will be explained in this section.

Errors in measurement systems can be divided in two parts, *systematic errors* and *random errors*. Systematic errors are those measurement errors that in replicate measurements remain constant or vary in a predictable, deterministic manner. Random errors are those measurement errors that in replicate measurements vary in an unpredictable manner.

As an example of a predictable systematic error one can think of the influence of temperature on the measurement of another physical quantity. It is often possible to measure the temperature and compensate for its influence, when the physical model of the system regarding temperature is known. An example of a random error is for instance the momentary voltage value of the thermal noise in a resistor. In order to decrease the error in measurement systems it is obvious that it is necessary to especially reduce the amount of random errors to a minimum.

The *measurement uncertainty* of a measurement system is quite strictly defined as a non-negative parameter, characterising the dispersion of the quantity values, being attributed to a measurand, based on the information used. It includes all random and non-compensated systematic errors. In case of compensation of predictable systematic errors it also includes random errors in the measurement that is used for the compensation, like the temperature measurement as mentioned in the example before.

The term *measurement accuracy* is a relative, mainly qualitative term, mostly used when indicating the difference between two measurements of which one is more accurate than the other. When quantised the given numbers are often only relating to the order of magnitude and as such this term is less strict than measurement uncertainty.

The term *measurement precision* is a relative, quantitative term, usually

expressed numerically by measures of imprecision relative to the maximum value of the measurand. One can think of the standard deviation, variance, or coefficient of variation under the specified conditions of measurement. In fact, measurement precision relates measurement uncertainty to the real value.

Last but not least, the term *measurement resolution* refers to the smallest change in a quantity being measured that causes a perceptible change in the corresponding indication. It is very important to recognise that the maximum resolution of a measurement system is quite something different than uncertainty. In fact the resolution has always the smallest value of the two, for which reason several suppliers of measuring systems like to only mention the resolution in their marketing communication.

### 8.1.1.1 The ultimate in uncertainty

Many years ago scientists like Laplace were convinced that somewhere in the future all physical processes would be exactly predictable.

Quantum mechanics however, has given us other insights with for instance the *Heisenberg uncertainty principle*, postulated by the German physicist and Nobel prize winner Werner Karl Heisenberg (1901 – 1976). His principle states that the standard deviation in the measurement errors of two parameters, like place ( $\sigma_x$ ) and impulse ( $\sigma_p$ ) or energy ( $\sigma_E$ ) and time ( $\sigma_t$ ), can not be smaller in combination, than given by the following expression:

$$\sigma_x \sigma_p \geq \frac{h}{4\pi} \quad \text{and} \quad \sigma_x \sigma_p \geq \frac{h}{4\pi}$$

with the Planck constant  $h = (6.6260693 \pm 0.0000011) \cdot 10^{-34}$  [Js].

Fortunately in the practical world of mechatronic reality this number is so small that it is not necessary to take this uncertainty limit into account. There is still ample room for scientists and engineers to work on improvements in the precision of measurement systems and keep consistently shifting the limits in measurement uncertainty.

### 8.1.2 Functional model of a measurement system element

Every element of a measurement system can be modelled as a separate subsystem with a primary input for the useful information, secondary inputs of external error sources and an output. This model is represented in Figure 8.2.

Ideally an element would only show a proportional linear transfer function  $L(i)$  from the measurand input  $i$  to the output  $o$ , with a transfer gain of  $K_\ell$ . In most sensors however a certain level of non-linearity causes a not fully proportional output value. This non-linearity can be expressed mathematically as a series-expanded polynomial that contains the higher-order terms of the total transfer function of the element.

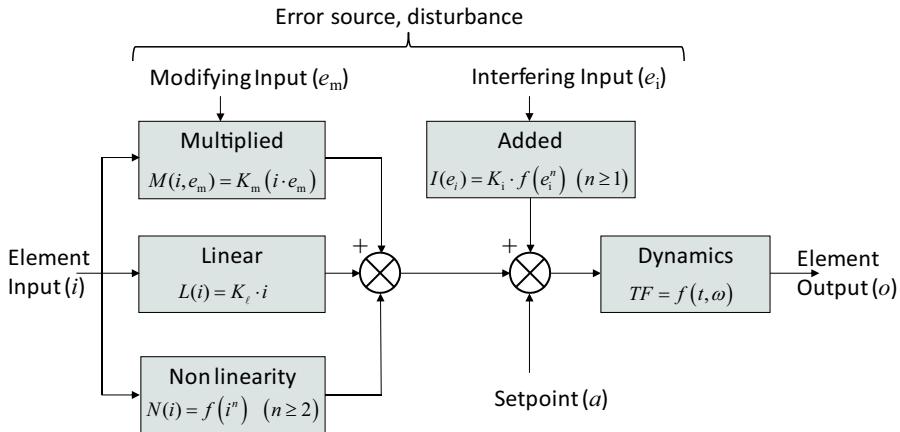
$$N(i) = \sum_{n=2}^{\infty} c_n i^n \quad (8.1)$$

Mostly the used terms are limited to only those that contribute to the significance of the output value  $o$ , as determined by the uncertainty level of the measurement.

Next to the non-linearity, external influences can impair the functionality of the element in two ways. Firstly, the external disturbance input can result in an additional value of the output of the element irrespective of the measurand input value. This disturbance input is called the *interfering input*, with a transfer gain  $K_i$ . The second way of impairing the measurement output is by changing the proportional gain of the measurand input transfer function and is called the *modifying input*, with a transfer gain  $K_m$ . In principle, the interfering input can have a similar non-linear transfer function as the measurand input, but often the higher-order terms of the related polynomial can be neglected, as the sensitivity of the interfering input is mostly much less than the sensitivity of the measurand. In high precision systems, however, this linearisation is not always allowed, especially when these effects need to be compensated.

Many external influences show a combination of both effects. Sometimes even the gain factors  $K_m$  and  $K_i$  depend on the actual value of the input or output signal, which makes it even more awkward to model. After compensation of the systematic errors, a precision system generally mainly shows random interfering errors.

Two other factors that influence the output of the element are shown in the model, the setpoint and the dynamics. The setpoint ( $a$ ) is an intentional offset-value to define the range of output values that belongs to a certain



**Figure 8.2:** Functional model of an element in a measurement system. In case of a sensor the element input is the measurand. Two ways of disturbing the output signal are shown. The modifying input changes the gain of the transfer function of the element input, while the interfering input just adds a signal to the output. Also non-linearity and dynamics can influence the functionality of the element.

range of input values. For instance with a thermal sensor, a temperature range of  $5 - 25^\circ \text{ C}$  could correspond with an output voltage range of  $0 - 10 \text{ V}$ . This example would require a linear transfer gain  $K_\ell = 0.5$  and a setpoint of  $a = 5 \text{ V}$ .

The Dynamics box represents the time or frequency dependent behaviour of the element and can be modelled with dynamic transfer functions. A typical example of an element with a first-order dynamic transfer function is a thermal sensor. The dynamic behaviour is determined by the heat capacity of the sensor and the heat conductivity between the sensor and the body to be measured. Second- and higher-order transfer functions are most often present in mechatronic positioning systems when the position sensor is located at a different place than the actuator.

## 8.2 Dynamic error budgeting

Random errors can only be quantified by means of statistics because of their unpredictability. Precision measurements are often repeated several times to reduce the impact of random measurement errors by applying statistical methods. Measurements in mechatronic systems are mostly continuous and the errors are a combination of a multitude of continuous interfering random signals. Also these can be treated in a comparable statistical way. After a short introduction in the statistics of errors with repeated single measurements, this section concentrates on the statistics of signals by introducing the concept<sup>2</sup> of *Dynamic Error Budgeting* (DEB). This methodology to derive the total error of a system from its different dynamic contributors, has proven to be extremely valuable in determining and solving error sources that act on real measurement and positioning systems.

### 8.2.1 Error statistics in repeated measurements

In official traceable metrology, errors are investigated by means of the analysis of many measurements. The observed distribution of the different values gives information about the origin of possible error sources, their character and methods of improvement. The related field of probability statistics is extremely wide with many different distributions that are represented with a *Probability Density Function*  $p(x)$  (PDF), giving the probability  $p_{x_1, x_2}$  that a measurement value is observed within a certain range of values from  $x_1$  to  $x_2$ .

$$p_{x_1, x_2} = \int_{x_1}^{x_2} p(x) dx \quad (8.2)$$

The probability that the measurement can have any value, is by definition equal to one. This implies that the integral from  $-\infty$  to  $+\infty$  of  $p(x)$  always equals one. The related *Cumulative Probability Function* (CPF) is defined by the integration of  $p(x)$  from  $-\infty$  to  $x$ . Often the PDF and the CPF are represented in a graphical way, like shown in the example of the following section. This representation clearly illustrates these statistical properties of the system. On the horizontal axis the range of values is shown while the

---

<sup>2</sup>This section is derived from the Phd thesis of Leon Jabben from our laboratory in Delft. His thesis can be downloaded for more details.

vertical axis denotes the probability density or the cumulative probability. Generally both axes are linear scaled.

### 8.2.2 The normal distribution

In precision positioning systems, errors are observed as deviations in signals with corresponding frequencies and amplitudes. For a precision measurement system in general, all systematic errors are compensated and calibrated. This means that only uncorrelated random signals remain as error source. As mentioned these errors have are mainly an interfering character which means that they are independent of the measurand.

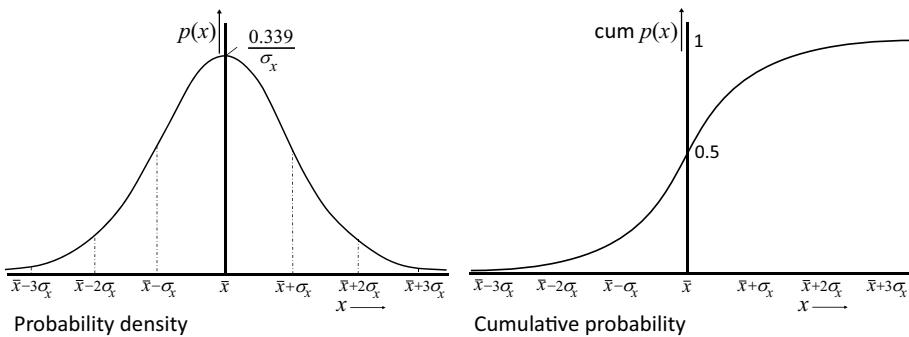
Random errors in signals can almost always be represented by means of a special probability density function, the *normal distribution* or *Gaussian distribution*, named after the same Johann Carl Friedrich Gauss, who postulated two of the Maxwell equations. The normal distribution is characterised by its mean value  $\bar{x}$  and its variance  $\sigma_x^2$ , while the square root of the variance is called the *standard deviation*  $\sigma_x$ . These terms are defined in the following way for multiple measurements:

$$\begin{aligned} \text{mean : } & \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \text{variance : } & \sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \end{aligned} \quad (8.3)$$

For random signals also a third term is introduced, the signal power  $P_s$  that proves to be useful in combining several error signals. For random signals the three terms are calculated as follows:

$$\begin{aligned} \text{mean : } & \bar{x} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t) dt \\ \text{power : } & P_s = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)^2 dt \\ \text{variance : } & \sigma_x^2 = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T (x(t) - \bar{x})^2 dt \end{aligned} \quad (8.4)$$

These definitions are somewhat extreme as in practice the “lim” term is never really necessary. Most disturbance signals are quite continuous and in practice a period  $T$  is chosen that sufficiently represents these interfering signals. A practical “rule of thumb” is to take  $T$  equal to the period of the lowest relevant frequency of the spectrum in the disturbance signal. For



**Figure 8.3:** The probability density function and cumulative probability function of a normal distributed random signal with a mean value  $\bar{x}$  and a standard deviation  $\sigma_x$ . The probability that a measurement value is located in a certain range is equal to the surface enclosed by the graph of the PDF and the horizontal lines that correspond with the range boundaries.

this reason in practice the “lim” term can be omitted.

It is also useful to remark that the variance equals the power and the standard deviation equals the root of the power for signals with a zero mean value. This standard deviation is then equal to the Root Mean Square value (RMS) as defined in Section 2.1.3.3 of Chapter 2.

Figure 8.3 shows a graphical representation of the probability density function and the cumulative probability function of a normal distributed random signal.

With the above definitions the mathematical expression of the probability density function is as follows:

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma_x^2}} \quad (8.5)$$

The calculation of the probability that the error is between  $\bar{x} \pm \sigma_x$ ,  $\bar{x} \pm 2\sigma_x$  or  $\bar{x} \pm 3\sigma_x$ , results in the following values:

$$\begin{aligned} p_{-\sigma_x, \sigma_x} &= \int_{-\sigma_x}^{\sigma_x} p(x) dx = 0.683 \\ p_{-2\sigma_x, 2\sigma_x} &= \int_{-2\sigma_x}^{2\sigma_x} p(x) dx = 0.955 \\ p_{-3\sigma_x, 3\sigma_x} &= \int_{-3\sigma_x}^{3\sigma_x} p(x) dx = 0.997 \end{aligned} \quad (8.6)$$

These calculations imply that only 68 % of the measured values can be found in a range of  $\pm 1\sigma$  around the mean value. A range of  $\pm 2\sigma$  already contains

more than 95 % of all measurements, while at  $\pm 3\sigma$ , less than 0.3 % is found outside the range. For this reason measurement systems are often specified with either  $2\sigma$  or  $3\sigma$  values for the measurement error.

### 8.2.3 Combining different error sources

Under certain conditions, different error signals can be combined by quite simple relations. In case of several mutually independent variables, it is allowed to combine the mean value and standard deviation of each variable by means of the method of the *root of the sum of squares*. When  $\alpha_n$  equals the proportion of the specific variable in the total measurement value, the following expressions are used to determine the mean value and the standard variation of the combined measurement:

$$\bar{x}_{\text{tot}} = \alpha_1 \bar{x}_1 + \alpha_2 \bar{x}_2 + \alpha_3 \bar{x}_3 + \dots$$

$$\sigma_{\text{tot}} = \sqrt{(\alpha_1 \sigma_1)^2 + (\alpha_2 \sigma_2)^2 + (\alpha_3 \sigma_3)^2 + \dots} \quad (8.7)$$

In practice it is often allowed to linearise the different effects because of the small variations of the inputs around a certain average value. This allows the use of simplified calculations with systems of which the errors are given in terms of their mean value and standard deviation.

The linearisation starts with the behaviour of an element according to the model written as a combination of the different inputs:

$$o = K_\ell i + a + N(i) + K_m i e_m + K_i e_i \quad (8.8)$$

For small deviations the non-linear term  $N(i)$  can be combined with the linear term  $L(i)$  and the equation becomes:

$$\delta o = \frac{\partial o}{\partial i} \delta i + \frac{\partial o}{\partial e_m} \delta e_m + \frac{\partial o}{\partial e_i} \delta e_i \quad (8.9)$$

where:

$$\frac{\partial o}{\partial i} \approx K_\ell, \quad \frac{\partial o}{\partial e_m} \approx K_m i \quad \text{and} \quad \frac{\partial o}{\partial e_i} \approx K_i \quad (8.10)$$

In most cases the different inputs are independent and their impact can be statistically combined using the rule of “the root of the sum of squares”. The combined standard deviation of the element becomes:

$$\sigma_o = \sqrt{\left(\frac{\partial o}{\partial i}\right)^2 \sigma_i^2 + \left(\frac{\partial o}{\partial e_m}\right)^2 \sigma_{e_m}^2 + \left(\frac{\partial o}{\partial e_i}\right)^2 \sigma_{e_i}^2} \quad (8.11)$$

And the median can be derived by simple addition:

$$\bar{o} = K_\ell \bar{i} + a + \overline{N(i)} + K_m \overline{i e_m} + K_i \overline{e_i} \quad (8.12)$$

### 8.2.4 Power spectral density and cumulative power

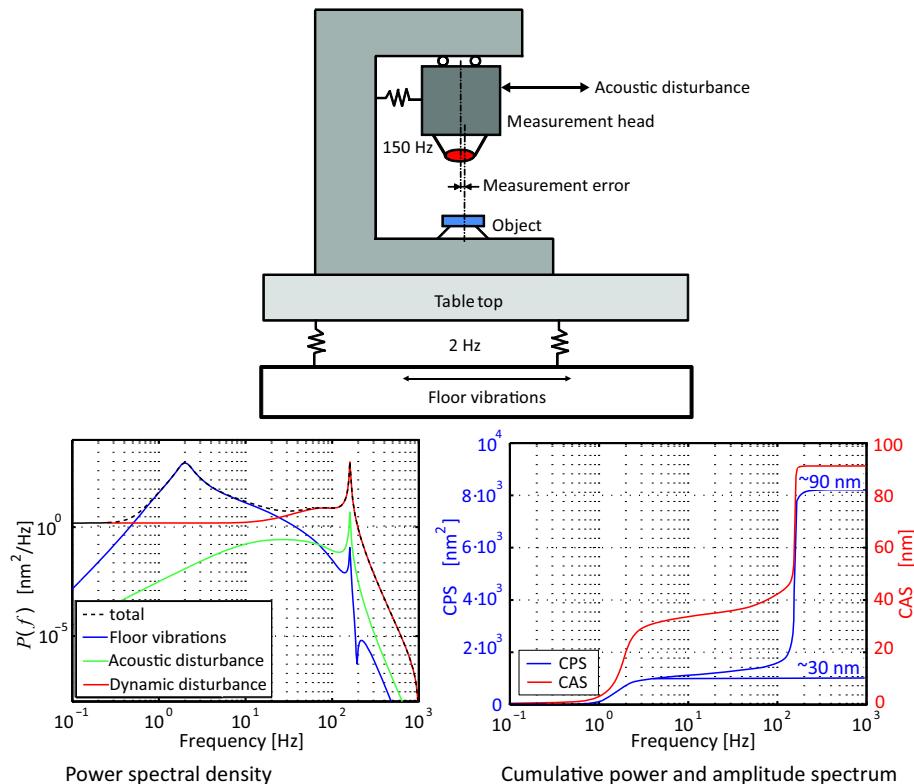
Dynamic error budgeting is a method to derive the total error in a dynamic system by using the described statistical methods to combine the contributions of all disturbance sources into one error value. The method starts with “the root of the sum of squares” that directly relates to the power of the different disturbance signals that act on a measurement system. When the total error of a measuring instrument by different error sources needs to be determined with this method, it is most useful to use two important related functions, the *Power Spectral Density* (PSD) and the *Cumulative Power Spectrum* (CPS). The power spectral density  $P(f)$  gives the power of a signal within a certain frequency range:

$$P_{f_1, f_2} = \int_{f_1}^{f_2} P(f) \quad (8.13)$$

The cumulative power spectrum is calculated by integrating  $P(f)$  starting from 0 Hz with increasing frequency until no further increase is observed. Also these spectra are shown in a graph with the frequency on the horizontal axis and the magnitude of the power density or the cumulative power on the vertical axis. The axes of the power spectral density are preferably double logarithmic, because the error signal characteristics often show a strong similarity with the Bode-plot of the dynamic transfer functions. The magnitude scale of the cumulative power spectrum is noted on a linear scale in order to avoid confusion with the contributions of the different error sources.

To illustrate this method of dynamic error budgeting, Figure 8.4 shows as an example the effect of three error sources on the inspection microscope that was introduced in Chapter 3.

The first error source is due to the transmission of floor vibrations to the sensitive part of the instrument. To prevent this transmission, a vibration isolation system is applied, consisting of springs that connect the mass of the table and the microscope with the floor. This vibration isolation system has a first eigenfrequency at 2 Hz. The shown power spectral density of this first error source is the combination of the unfiltered vibrations of the floor, the transmissibility transfer function of the vibration isolation system and the sensitivity of the instrument to this interfering error source. At low frequencies the microscope acts as a rigid body and the impact is low. At the resonance of 2 Hz a maximum is reached in the spectral density, with a negative slope at higher frequencies due to the transmissibility transfer function of the vibration isolation system. At 160 Hz the resonance of



**Figure 8.4:** The power spectral density, the cumulative power spectrum and the cumulative amplitude spectrum of an inspection microscope with different error sources, the floor vibrations, acoustic disturbance forces acting on the measurement head and the dynamic disturbance from the measurement head itself. The resonance of the vibration isolation system at 2 Hz and the resonance of the microscope head at 160 Hz show to be the main error contributors.

(Courtesy of Leon Jabben)

the measurement head shows the typical dynamic decoupling effect of this eigenmode.

The second error source originates from acoustical disturbances by for instance the air conditioning unit of the room, where the microscope is used. The power spectral density of the combination of these disturbances with the sensitivity of the instrument shows a maximum at the resonance frequency of 160 Hz.

The third error source is caused by the dynamic properties of a feedback controlled positioning system that moves the measurement head within

the microscope. By a bad design of the controller and a noisy sensor it increases the internal resonance of the microscope head which means that these dynamics are also most disturbing at the 160 Hz resonance frequency. The total power spectral density is the sum of all PSD graphs and due to the logarithmic scale, it is equal to the enveloping curve of the three error contributors. In this example the total error spectrum appears to be only determined by the floor vibrations and the dynamic disturbances by the positioning system of the microscope head. The acoustic disturbance effect remains below the error floor over the full frequency range.

The next step that is necessary to derive the total error is the determination of the cumulative power spectrum. Starting at low frequencies, a first rise in the cumulative spectrum is observed in the graph at 2 Hz and a much larger second rise at 160 Hz with a final power level of  $\approx 8.1 \cdot 10^3 \text{ nm}^2$ . This difference in magnitude is confusing at first sight, as the surface of the peak in the PSD at 2 Hz is much larger than the surface of the peak at 160 Hz. This is however caused by the logarithmic frequency scale of the PSD. Integrating the PSD over that low-frequency area only results in a relatively small cumulative value.

To determine the standard deviation of the final error, the root of the total CPS value should be taken, which is 90 nm in the above example. Due to the two eigenmodes in the system, this value is much higher than the 10 nm error that was calculated in Chapter 3. This difference is caused by the two undamped resonances and the largest contributor to this error is the badly designed controller of the positioning system of the microscope head.

As was shown in Chapter 4 on motion control, an undamped mass-spring system can be controlled with a well tuned PD-control setting such that the resonance is completely suppressed by shifting the poles to the left in the complex plane. In that case no resonance will occur in the dynamic disturbance and the step in the CPS at 160 Hz will become negligible. The remaining error of around 30 nm is then only determined by the floor vibrations. As a consequence the next improvement measures should be focused on the vibration isolation system where some damping might be sufficient to sufficiently reduce the peak at 2 Hz. Too much damping will however increase the disturbance by the floor vibrations at higher frequencies so this measure should be taken with precaution.

### 8.2.5 Cumulative amplitude

When trying to achieve the minimum error level in a complete measuring system, it is frequently observed that designers directly convert the CPS

graph into a *Cumulative Amplitude Spectrum* (CAS)<sup>3</sup> graph by deriving the root of the CPS values. Although the resulting fully cumulated end value remains the same, this representation is not advisable for problem solving. By taking the root, the impact of the disturbances at the low-frequency side is visually enlarged, relative to the impact of the disturbances at higher frequencies.

This is clearly illustrated by the red line of the CAS graph as compared with the blue line of the CPS graph. When observing the CAS graph only, a designer could erroneously conclude that the total error can be reduced with approximately 30 nm by simply adding damping to the vibration isolation system. Unfortunately however, the PSD would then only be reduced from  $\approx 8.1 \cdot 10^3 \text{ nm}^2$  to  $\approx 7 \cdot 10^3 \text{ nm}^2$ , resulting in an error amplitude of 83 nm, only an almost negligible 9 nm below the previous value. This counter intuitive effect is fully caused by "the root of the sum of squares".

### 8.2.5.1 Variations on dynamic error budgeting

The cumulative power spectrum can in principle also be determined by starting the integration at higher frequencies and integrating towards lower frequencies. This would result in a cumulative power spectrum with the same rise magnitudes at critical frequencies as when starting integration at low frequencies, so it would lead to the same conclusions. Based on such a reversed PSD a cumulative amplitude spectrum would however look completely different. In the example case of the inspection microscope the resonance at 160 Hz would be even more emphasised and the effect at 2 Hz would almost disappear. Also this other approach shows that only the cumulative power spectrum should be used for evaluating the impact of different error sources.

## 8.2.6 Sources of noise and disturbances

The sources of disturbing signals in mechatronic systems all have either a mechanical or an electronic nature while many of the observed phenomena have a thermal root cause. Large scale deformations by heating can partly be avoided by a careful design, that takes all expansion effects into account, but ultimately always some thermal effects will remain, especially when

<sup>3</sup>It should be noted that the word "Amplitude" in the CAS is erroneous as not the amplitude of a signal is calculated, but the standard deviation or the RMS value in the case that the mean value is zero.

the heat flow is not constant. Most probably this direct thermal disturbance on system accuracy will always remain a limiting factor in the precision of high performance mechatronics.

Thermal effects have their dynamic impact also on the small scale as they are the root cause for the noise in electronic components like resistors.

The large scale deformations by thermal effects mainly occur in the low-frequency area. In the following a short overview is given of some important sources of higher frequency dynamic noise in the mechanical and electronic domain:

#### 8.2.6.1 Mechanical noise

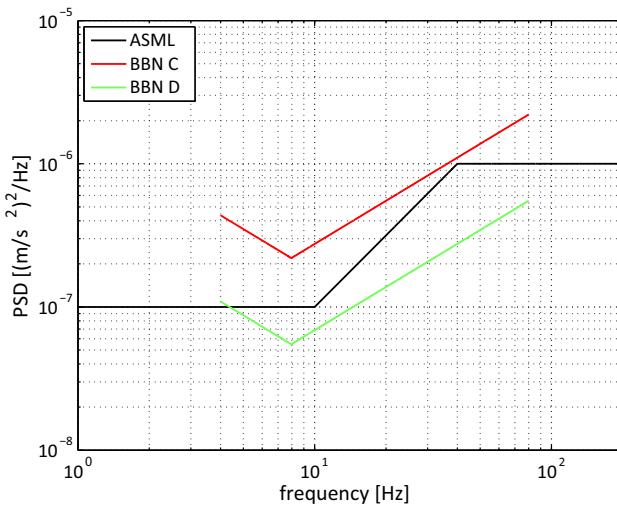
The most important source of mechanical dynamic disturbances are vibrations that either are caused by movement of the support, the floor vibrations, or by forces from vibrating parts in the machine itself. Floor vibrations can be caused by traffic, wind, earthquakes and other machines or moving people in the neighbourhood. The power spectral density of floor vibrations can consist of real uncorrelated random signal and systematic signals caused for instance by rotating equipment. When the ultimate of precision is aimed for, it is always necessary to analyse the vibrations on a certain location with measurements over a representative time span, at least over a full twenty-four hours period on a busy day. For production equipment like wafer scanners it is not possible to adapt the machine to all different locations and it is better to work with a certain standard specification. Figure 8.5. shows the specification of an ASML wafer scanner in comparison with the *BBN criteria* that were defined by the American company Raytheon BBN Technologies.

#### 8.2.6.2 Electronic noise

Electronic noise has many origins, which are named after their behaviour, like thermal, shot, excess, burst and avalanche noise. The most relevant sources of electric noise that impair precision mechatronic systems is presented in the following short overview:

##### Thermal noise

Any resistor will have a fluctuating potential difference across its ends that is superimposed on the voltage caused by the current through the resistor.



**Figure 8.5:** Power spectral density of floor vibration specifications from ASML and BBN.  
(courtesy of Leon Jabben)

The fluctuating voltage is caused by the thermally induced random motion of charge carriers like electrons. Thermal noise has a normal probability density function and it has a flat power (*white noise*) spectral density, called. In electric systems the energy is dissipated in the resistors. The noise from a resistor can be described as a voltage source in series with the resistor, with a power spectral density of:

$$N_T = 4kTR \quad [\text{V}^2/\text{Hz}], \quad (8.14)$$

with  $k$  the Boltzmann's constant ( $1.38 \cdot 10^{-23}$  J/K),  $T$  the temperature and  $R$  the resistance. To give an example, a resistor of  $1 \text{ k}\Omega$ , at  $20^\circ\text{C}$  will show noise with a RMS value of  $0.13 \mu\text{V}$  from 0 Hz up to 1 kHz.

### Shot noise

Shot noise results from the random passage of individual charge carriers across a potential barrier. This is often seen with junctions in a transistor. The noise has a normal probability density function and has a white spectral density:

$$N_S = 2qIDC \quad [\text{A}^2/\text{Hz}], \quad (8.15)$$

with  $q_e$  the charge of an electron ( $1.6 \cdot 10^{-19}$  [C]),  $IDC$  the average current [A]. An average current of 1 A will introduce noise with an RMS value of 18 nA from zero up to one kHz.

### Excess noise

The noise in excess of the thermal and shot noise when a current passes through a resistor or a semiconductor, is called excess noise. Other names are *flicker noise* or *one-over-f (1/f) noise*. This noise source results from fluctuating conductivity due to imperfect contact between two materials. This is the reason why carbon composition resistors, which are made up of many tiny particles molded together, show more excess noise than wire wound resistors. The power spectral density of excess noise increases when the frequency decreases:

$$N_E = \frac{K_f}{f^\alpha} \quad [\text{V}^2/\text{Hz}], \quad (8.16)$$

where  $K_f$  is dependent on the average (DC) voltage drop over the resistor and the index  $\alpha$  is usually between 0.8 and 1.4, and often set to unity for an approximate calculation. For resistors the excess noise is proportional to the average voltage drop  $V$  over the resistor, which is why manufacturers typically specify the excess noise as a noise index  $C_R$  for one frequency decade:

$$C_R = \frac{\sigma_V^* \cdot 10^6}{V} \quad [\mu\text{V}/\text{V}] \quad (8.17)$$

with  $\sigma_V^*$  being the standard deviation over one decade frequency range of the voltage. For standard resistors the noise index  $C_R$  typically ranges from 1 to 10. For example, if the noise index equals 10, an average voltage drop of one volt introduces noise with an RMS value of 17  $\mu\text{V}$  in a frequency range from 1 up to 1000 Hz. Note that a frequency range 1 mHz up to 1 Hz introduces an equal amount of noise because of the low-frequency range!

#### 8.2.6.3 Using noise data from data-sheets

The given data on noise in different data-sheets of electronic components are most frequently given in terms of  $[\text{V}/\sqrt{\text{Hz}}]$  as an RMS density instead of the power density. This is done to be able to easily calculate the noise-voltage contribution of one single electronic component. It is important to emphasise here that these values should first be squared and then combined

to determine the total power spectral density of the system.

Many errors have been made in the past by designers who either used the RMS data in a PSD and got a very low disturbance level after taking the root of the CPS. But also a non-statistical worst-case addition of the noise voltages per electronic element gives erroneous results as the noise voltages are mostly uncorrelated and as a consequence the worst-case addition gives a too high value.

These errors either result in an unobserved problem that will pop-up in the realisation of the mechatronic system when the real measurements will unveil the noise or it will give rise to a too expensive design, because the calculated worst-case noise value had to remain below the specification of the total system.

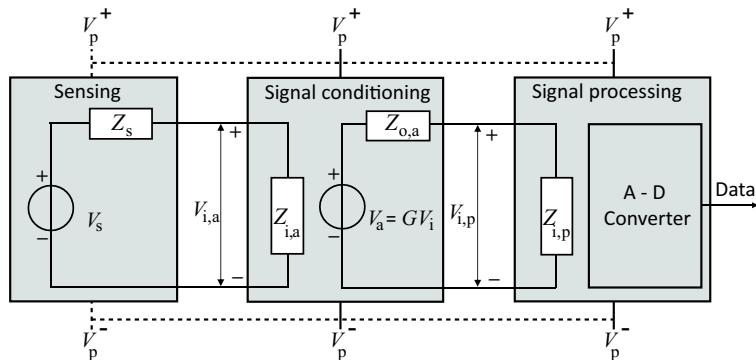
In any case these errors can be avoided by using the method of Dynamic Error Budgeting in the right way.

### 8.3 Sensitive signals in measurement systems

Most of the information exchange inside and between all elements in a measurement system takes place in the electronic domain. Figure 8.6 shows a simplified electronic model of the measurement system. The input impedance of every successive element determines the load for the output of the preceding element. Depending on the different properties of each element the requirements for the successive element can be quite extreme. In most cases the output impedance of the sensing element is reasonably high. A large output current with a high voltage would correspond with a high power level that has to originate from the physical phenomenon that is measured. This means that in general the input impedance of the successive element should be either be as high as possible with a very small input current or as low as possible with a very small input voltage. In both cases the input power will be limited.

Also the power supply connections are shown. The power supply of the sensing element is dashed because not all sensing principles need an external power input. Still all these connections can and will cause errors of a mainly interfering character, when not designed well.

This section will deal with the most sensitive part of the measurement chain, the sensing element and its interconnections.



**Figure 8.6:** The electronic model of a measurement system shows the connection between the three elements and the power supplies. Every successive element determines a load for the preceding element. The output of the first two elements is shown in the Thevenin model with a voltage source. Depending on the properties of the system also a current source Norton equivalent model can be applied.

### 8.3.1 Sensing element

The sensitivity of a sensing element for a certain physical quantity is based on the effect of that quantity on an electrical property of that sensor. Almost by definition this does not exclude sensitivity for other physical phenomena that will act either as an interfering input, as a modifying input or as a combination of both. This is the first problem that a mechatronic designer encounters, when choosing a suitable sensor. The most prominent example of this phenomenon is the influence of temperature on almost every electrical property of any sensor. Even when measuring the temperature itself, the measurement is often influenced by temperature values elsewhere in the measurement system.

The following non exhaustive list shows some representative sensing principles. They can be distinguished into two different operation principles. The first principle relates to the capability of a sensor to directly generate an electrical signal by converting energy from the physical phenomenon into electric energy. The second principle is based on the variability of an electrical impedance by the physical phenomenon.

Examples of direct generation of an electrical signal are:

- The voltage difference between different metals to measure temperature differences.
- The voltage induced in a coil by a changing magnetic field, proportional to the relative velocity.
- The voltage induced over a piezoelectric crystal by deformation due to a force.
- Electrons in the depletion layer of a diode that are excited by photons of the incident light and create an electric current.

Examples of a variable electrical impedance are:

- The change of resistance of a resistor by temperature or strain.
- The change of the capacitance value of a capacitor by a displacement of the electrodes.
- The change of the self inductance of an inductor by a change in the magnetic geometry.

Not all sensors are capable of directly converting a physical quantity into electricity. They first need to create one of the primary effects from the list

above by additional preceding steps. Take for example the measurement of Force. Force is a physical quantity that can not be directly measured as it manifests itself either as the acceleration of a mass or by deformation of material. This means that for measuring a force, it is necessary to measure either the corresponding acceleration or the deformation. Later in this chapter it will be shown that acceleration is in most cases measured by measuring the deformation of a material that is caused by the force that is needed to accelerate a known seismic mass.

### 8.3.2 Converting an impedance into an electric signal

In the previous section it was shown that sensing principles can be based on a changing electrical impedance. Before this information can be treated in the same way as a directly generated electrical signal, it is first necessary to convert this impedance into an electrical signal<sup>4</sup>.

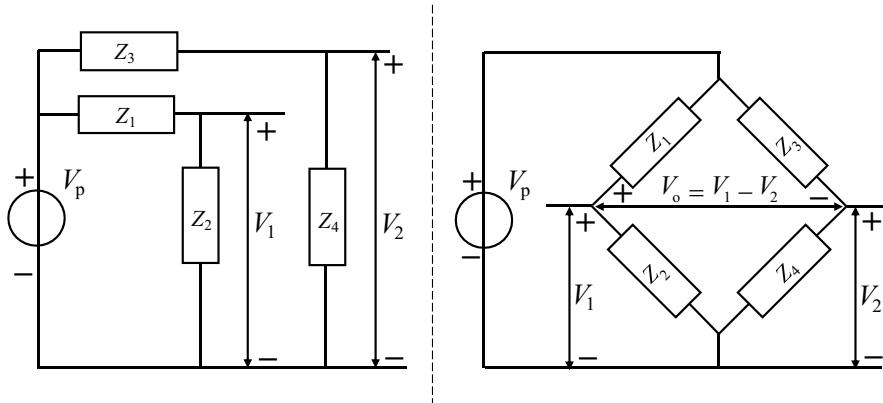
The most simple way to achieve this conversion is by applying Ohms Law and either supply the impedance with a well-known current and measure the resulting voltage over the impedance or do just the opposite by applying a voltage source and measuring the current. An AC source must be used with a complex impedance like a capacitor or an inductor, but apart from that the reasoning is the same.

Although modern electronics can be made rather precise, often special measures are needed to prevent errors that are related to the low sensitivity of many sensing impedances. When a potentiometer is used to measure the displacement of the slider, the output voltage ranges from the supply voltage  $V_i$  to zero. This is a sufficiently large signal, when compared with interfering signals, and generally no further measures are needed. Another example of a very sensitive resistive sensor is the *NTC*, a resistor with a negative temperature coefficient based on semiconductor properties. Its resistance can for instance range between 50 and 1000  $\Omega$  for a temperature change from 0 – 100 °C. With 1 mA current this results in a sensitivity of approximately 10 mV/°C, which is also sufficient for most practical measurements.

Unfortunately most variable impedance sensors have a much lower sensitivity and the noise in the supply current or voltage will play a much larger role with those sensors. Take for instance a *strain gage*, a resistor to measure the strain in a material by an increase of its resistance, when elongated. With

---

<sup>4</sup>Often this conversion is considered as a part of the signal conditioning element. This is however not logical with the aforementioned definition of the sensing element



**Figure 8.7:** A Wheatstone bridge has two branches each consisting of a voltage divider. Only the differential voltage between the outputs of both branches is used as measurement signal. Frequently the Wheatstone bridge is drawn in a tilted square configuration as shown right to emphasise the differential output voltage  $V_o$ . When all resistors are equal the differential voltage is zero and the common-mode voltage is half the supply voltage.

those elements, the practical resistance change upon load  $\delta R/R$  is often less than 1 %. When this strain gage is supplied with a constant current source  $I_p$ , it is very difficult to keep the noise in this source below  $10^{-6} \times I_p$ . With an average value of  $R = 100 \Omega$ , this noise causes an error with a constant RMS value in the measurement of  $10^{-6} \times 100 \Omega$ . With  $\delta R/R \leq 0.01$ , the signal to noise ratio is less than  $10^4$  and because the error signal has a constant RMS value this is unacceptable for any precision measurement system. Also the temperature has a large effect on most resistive sensors and this influence directly interferes with the resistance change of the measurement.

### 8.3.2.1 Wheatstone bridge

To reduce the negative effects associated by the noise in the voltage source of impedance measurement electronics, sensors like strain gages are mostly connected in a *Wheatstone bridge*, named after the British scientist Sir Charles Wheatstone (1802 - 1875), who popularised its use after the invention by the British scientist Samuel Hunter Christie (1784 - 1865).

The thinking model behind the usefulness of the Wheatstone bridge is best explained by means of Figure 8.7. In principle the bridge consists of two voltage dividers that share the same voltage source. These voltage dividers

are also called the two *branches* of the Wheatstone bridge. The first branch consists of  $Z_1$  and  $Z_2$  with a corresponding output voltage  $V_1$ . Similarly the second branch consists of  $Z_3$  and  $Z_4$  with output voltage  $V_2$ . The total measurement signal is obtained by using an ideal differential amplifier to amplify only the difference voltage  $V_o = V_1 - V_2$  without loading the bridge. The benefit of this configuration becomes clear when starting in the situation that all four resistors are equal. This would result in equal voltages  $V_1 = V_2 = 1/2V_p$  and the voltage difference  $V_o$  would then be zero. This is called an ideally *balanced bridge* and in that case the influence of the power supply voltage  $V_p$  would be cancelled in the differential voltage. It would only create a common-mode noise voltage and this will be rejected by the differential amplifier.

As soon as one of the resistors, for instance  $Z_2$ , changes its value, it will cause  $V_1$  to change, resulting in a non-zero voltage difference  $V_o$ . In this new situation the noise of the voltage source will also be observed in the differential voltage signal but at a much smaller value, proportional to the measurement signal. With the same sensor and source voltage of the previous example, this would result in a relative RMS noise level of  $10^{-6}$  of the signal value, which is far better than the previously obtained constant RMS noise level of  $10^{-4}$  of the maximum signal value.

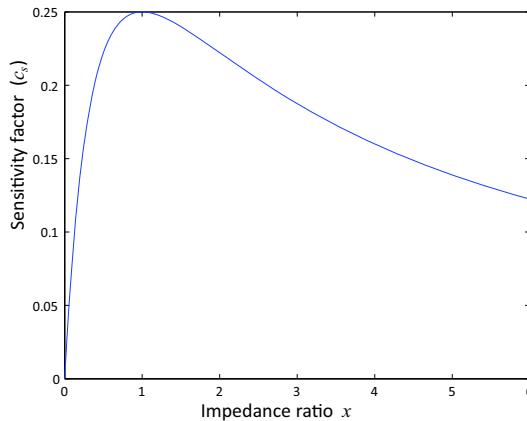
In theory it is not really necessary that all resistors are equal to realise a balanced bridge, as long as  $V_1$  remains almost equal to  $V_2$ . This condition is always met, when the ratios of both branches of the Wheatstone bridge are equal ( $Z_1 : Z_2 = Z_3 : Z_4$ ). For example one of the branches can consist of resistors with much higher values than those in the other branch, like  $Z_1 = Z_2 = 2 \text{ k}\Omega$  and  $Z_3 = Z_4 = 200 \Omega$ .

In principle also  $Z_1$  can be different from  $Z_2$  as long as their ratio is equal to the ratio between  $Z_3$  and  $Z_4$  but that is in most cases not a preferred situation. With most sensors the impedance variation is only a small fraction of the nominal impedance of the sensor, proportional to the measurand  $i$ , so  $\delta Z/Z \propto i \ll 1$ . For those sensors a maximum sensitivity is obtained when  $Z_1 = Z_2$  and  $Z_3 = Z_4$ .

This statement can be proven by calculating the sensitivity of for instance  $V_1$  as function of its measuring impedance  $Z_2 = R + \delta R$  for different values of the other impedance  $Z_1 = xR$ . With  $\delta R/R = c_m i$ , where  $c_m$  is a constant,  $\delta R$  can then also be written as  $R c_m i$  and the voltage  $V_1$  becomes:

$$V_1 = V_p \frac{R + \delta R}{xR + R + \delta R} = V_p \frac{R(1 + c_m i)}{R(x + 1 + c_m i)} = V_p \frac{1 + c_m i}{x + 1 + c_m i} \quad (8.18)$$

The sensitivity can be approximated as mainly linear because  $c_m i \ll 1$  and



**Figure 8.8:** The sensitivity factor of a wheatstone bridge is maximum when the impedance values in each branch are equal ( $x = 1$ ), resulting in a working point of half the supply voltage.

the corresponding value of the linear gain  $K_\ell$  becomes equal to the derivative of  $V_1$  over  $i$ :

$$\begin{aligned} K_\ell &= \frac{dV_1}{di} = V_p \frac{(x + 1 + c_m i)(c_m) - (1 + c_m i)(c_m)}{(x + 1 + c_m i)^2} \\ &= V_p \frac{xc_m}{(x + 1 + c_m i)^2} \approx V_p c_m \frac{x}{(x + 1)^2} = V_p c_m c_s \end{aligned} \quad (8.19)$$

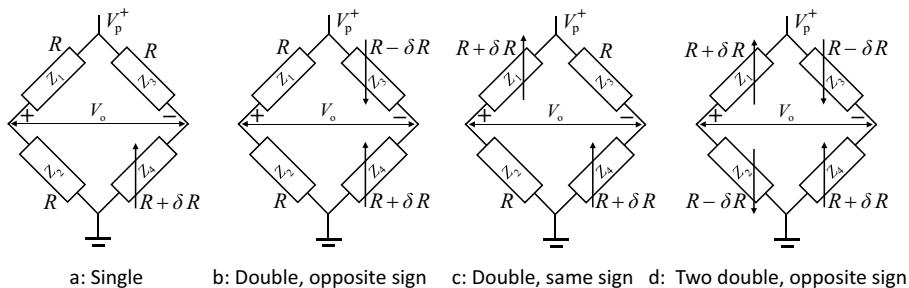
with  $c_s$  being the sensitivity factor of the Wheatstone bridge. The mentioned approximation is allowed when  $c_m i \ll 1$ . The maximum sensitivity is found when the derivative of  $c_s$  over  $x$  becomes zero:

$$\frac{dc_s}{dx} = \frac{(x + 1)^2 - x(2x + 2)}{(x + 1)^4} = \frac{(x + 1)^2 - 2x(x + 1)}{(x + 1)^4} = \frac{1 - x}{(x + 1)^3} \quad (8.20)$$

This derivative is zero for  $x = 1$  and  $x = \infty$ , but the second value of  $x$  also gives a zero value for the sensitivity  $K_\ell$ . This means that only  $x = 1$  is a useful answer, proving the statement that the values of the impedances in each branch of the Wheatstone bridge should be approximately equal for a maximum sensitivity.

This conclusion is further illustrated in Figure 8.8, emphasizing the non-linear relation between the sensitivity and the ratio of the impedances. Only in a small area around  $x = 1$  the sensitivity is almost constant.

As a last variation on the Wheatstone bridge, the two branches might differ in the kind of impedance that is applied. For instance one branch can consist of two resistors, while the other branch can consist of two complex



**Figure 8.9:** Four configurations of a Wheatstone bridge with maximum sensitivity, showing the different configurations that can be chosen when a multiple of variable impedances are available. Versions a: and c: suffer from a small non-linearity and both are not temperature compensated. The other two versions are both linear around the working point and temperature compensated.

impedances of the same kind, like two capacitors or two inductors. Other combinations with for instance only one capacitor are not allowed as then the output voltages would show a relative phase shift for AC voltages because of the filtering. This would result in a measurable voltage difference with a corresponding noticeable noise level, even without measurement signal. The application of complex impedances will be demonstrated more in detail in Section 8.6.2 and thereafter with the presentation of capacitive and inductive proximity sensors.

### Temperature compensation and linearisation

When all resistors in a balanced resistive Wheatstone bridge share the same temperature and the same temperature sensitivity, the output voltage is also no longer affected by the temperature. This seems attractive at first sight, but these conditions are not easily met, because the temperature of all elements are hardly ever equal. The not-sensing impedances of the bridge are in most cases located inside the measurement instrument, at some distance from the measurement location. In principle the requirement for an equal temperature can be alleviated a bit as only each branch needs to be isothermal. This is based on the reasoning that in that case the ratio of both branches is not affected by the temperature. This means that one of the branches can remain inside the measuring instrument at an equal temperature, while the other branch is fully located at the measurement site.

When possible, this second element in the sensor branch of a Wheatstone bridge should also be used for active measuring the phenomenon of interest. As will be demonstrated in the following this second sensor would be optimally applied, when it gives an opposite change of its impedance with an equal magnitude as the first sensor impedance. To investigate this effect, first the sensitivity of the full Wheatstone bridge is written down. The bridge is assumed optimal in respect to sensitivity so all impedances are as much as possible equal. Using the notation for the impedances from Figure 8.9, the output voltage  $V_o$  is given by the following generic equation:

$$V_o = V_p \left( \frac{Z_2}{Z_1 + Z_2} - \frac{Z_4}{Z_3 + Z_4} \right) \quad (8.21)$$

With the given values, the output voltage of the single sensor version with  $Z_4 = R + \delta R$  and  $Z_1 = Z_2 = Z_3 = R$  becomes:

$$V_o = V_p \left( \frac{1}{2} - \frac{R + \delta R}{2R + \delta R} \right) \quad (8.22)$$

After combining the terms with a common denominator, this equation can be written as:

$$V_o = -V_p \frac{\delta R}{4R + 2\delta R} \approx -V_p \frac{\delta R}{4R} \quad (8.23)$$

The approximation is allowed as long as  $\delta R \ll 2R$ . In the not approximated equation the  $\delta R$  term in the denominator causes a small non-linearity in the sensitivity. Also the thermal effects in  $Z_4$  are not compensated.

The situation becomes however quite different with two sensing elements with an opposite sign of the impedance change. This is shown in the second drawing of Figure 8.9, where the second sensing impedance is located at the other impedance of the same branch, so  $Z_3 = R - \delta R$ . With this double sensing principle the output voltage becomes:

$$V_o = V_p \left( \frac{1}{2} - \frac{R + \delta R}{2R} \right) = -V_p \frac{\delta R}{2R} \quad (8.24)$$

This equation is fully linear with a double sensitivity, when compared to the single sensing principle and also the thermal effects are compensated as both sensing elements are located at the measurement site.

It should be noted that this linearity is limited to the optimal working point with all resistors approximately equal.

At first glance the second sensor could also be located in the other branch at  $Z_2$ . With very small impedance changes this indeed would give approximately the same effect on the differential voltage but the non-linearity

would not be cancelled as can be seen by filling in the data:

$$V_o = V_p \left( \frac{R - \delta R}{2R - \delta R} - \frac{R + \delta R}{2R + \delta R} \right) = -V_p \frac{2R\delta R}{4R^2 - 4R\delta R + (\delta R)^2} \approx -V_p \frac{\delta R}{2R} \quad (8.25)$$

In this case the thermal effects are still compensated as these work equal in both branches but the non-linear terms in the not approximated equation still underline the optimal location of both sensors in the same branch.

The negative effect of using two sensors in different branches is also demonstrated in the third configuration of Figure 8.9. This configuration is unfortunately the only option, when the second sensing element has the same sign for its sensitivity as the first sensing element. In that case for instance  $Z_1$  and  $Z_4$  are replaced by the two equally acting sensing impedances  $R + \delta R$ . The calculation of the output voltage gives the following result for this situation:

$$V_o = V_p \frac{\delta R}{2R + \delta R} \approx V_p \frac{\delta R}{2R} \quad (8.26)$$

Although the sensitivity is also approximately twice the value of the single sensing principle, the non-linearity is not cancelled. The main disadvantage is however that the thermal problem is not solved, because of the different location of the impedances in their branch. A simultaneous equal change in the impedance by the temperature will cause a differential interfering voltage adding to the voltage of the useful measurement signal.

It should be noted that the same result but with a different sign would be obtained when  $Z_2$  and  $Z_3$  are replaced by the two sensing elements instead of  $Z_1$  and  $Z_4$ .

Thinking further on this path, it appears that the most optimal configuration is created, when all four impedances of the Wheatstone bridge would be replaced by sensing elements. This results in a total signal of:

$$V_o = V_p \frac{R - \delta R}{R + \delta R + R - \delta R} - \frac{R + \delta R}{R + \delta R + R - \delta R} = -V_p \frac{\delta R}{R} \quad (8.27)$$

In this configuration, both the temperature effects and the impact of the noise of the source is minimised, the sensitivity is maximum, the sensor is linear around the working point and only 4 connections are needed to the measurement electronics.

It will be shown in Section 8.7.1 how this can be accomplished in an integrated strain gage element.

**Remark on non-linearity:** The argumentation regarding non-linearity has become less important over the years with the introduction of digital

data processing. In principle non-linearity is a deterministic error and it can be compensated with the described model. It is certainly true to say that compensation is as good as the model is described and nothing is ideal so “prevention is still always better than curing”, but in many situations a less optimal bridge configuration can be compensated to an acceptable error level.

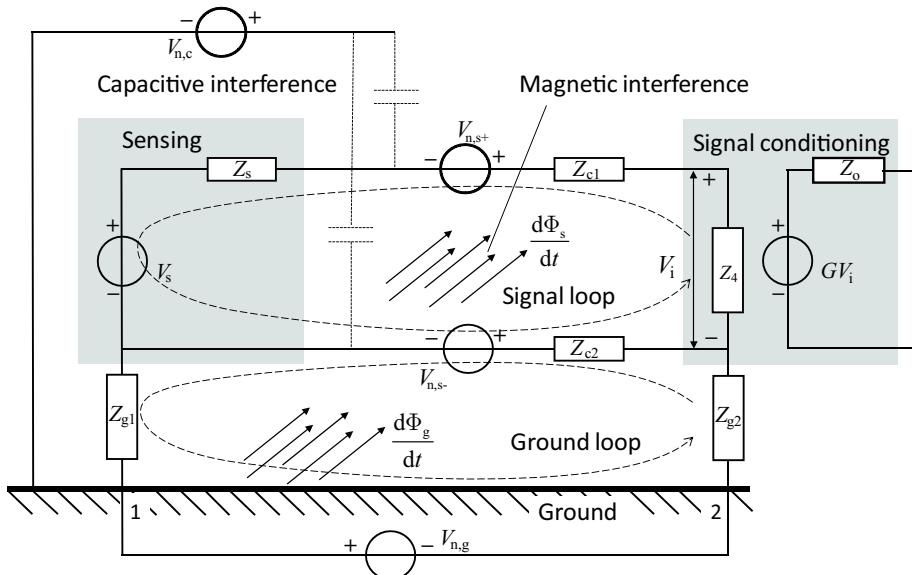
### 8.3.3 Electronic interconnection of sensitive signals

Like mentioned before, in most cases the electrical signals that are generated in a sensing element, are still susceptible to interfering signals from external sources of energy, because of their small amplitude and high source impedance. The sensitivity for these disturbances demands a careful consideration of the interconnection from the sensing element to the signal conditioning element. Although this interconnection is the most critical in a measurement system also the other connections can cause problems but first the purely analogue most sensitive interconnection will be examined. All elements have to be interconnected with minimal two wires, because electricity flows in a loop, and both wires are susceptible to interference. Figure 8.10 gives an overview of the interconnection between two elements of a measurement system with the different external disturbance sources that can act on this interconnection.

#### 8.3.3.1 Magnetic disturbances

The first source of disturbances is related to voltages that are induced in a closed electric loop by changing magnetic fields from power transformers and high currents in nearby electronic circuits. For both wires these voltages have mostly a different value. The reason is that their corresponding loop encloses a different surface, resulting in a voltage difference  $V_{n,s}^+ - V_{n,s}^-$  that is added to the output voltage of the first element. When these voltages would be equal, the differential voltage would be zero and only a common-mode disturbance input voltage would be present at the next element. This common-mode voltage was defined in Chapter 6 with the operational amplifier and represents the average value of both input voltages. This impact of a common-mode voltage can be reduced by using a differential amplifier with a high “common-mode rejection ratio” in the following element.

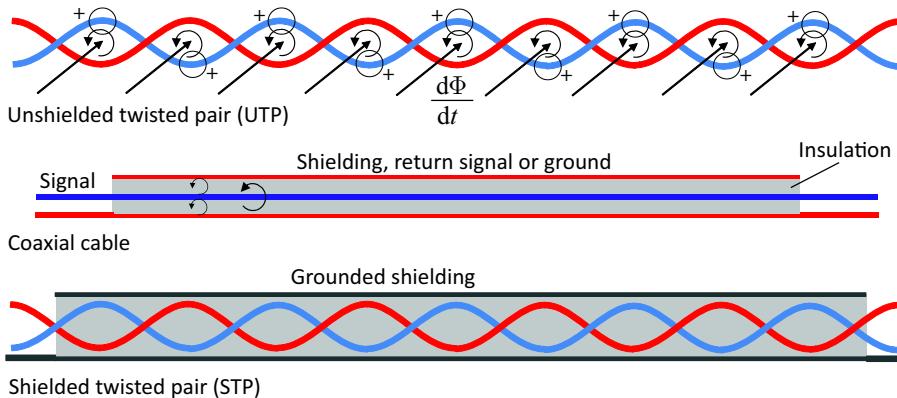
Unfortunately it is impossible to have the two wires follow exactly the same path as then they would make contact and create a short circuit. It is also



**Figure 8.10:** Overview of different sources of interfering signals acting on the connection between two elements in a measurement system. Changing electric fields insert a capacitive current in the connecting cables and changing magnetic fields induce a voltage in closed loops. Grounding at different places includes the voltage differences of these places in the measurement.

not possible to isolate a sensitive system from magnetic fields as no materials exist with a relative magnetic permeability  $\mu_r$  of zero that would be needed to lead magnetic fields away from the sensitive electronics. Magnetic fields can only be reduced to some extent by fully closed thick layers of ferromagnetic material for low-frequency magnetic fields and by a fully closed conductive shielding for high-frequency magnetic fields (eddy-currents), but that solution is very unpractical in many situations.

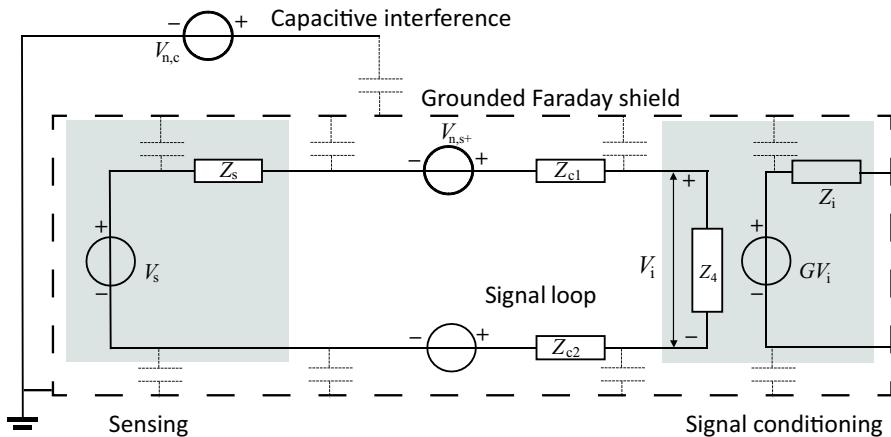
To solve this *magnetic interference* in a more practical way, two methods are used. For less critical situations the *Unshielded Twisted Pair* (UTP) cable can be used as shown in the upper drawing of Figure 8.11. A homogeneous changing magnetic field will induce an electromotive force in each loop that encloses a part of this field, according to Faraday's law. As long as every twist encloses the same magnetic flux an equal electromotive force is induced. Every twist is succeeded with a next twist where the wires are exchanged from position resulting in an exchanging positive and negative electromotive force in each twist for both wires. These electromotive forces



**Figure 8.11:** Interference by changing magnetic fields can be reduced by twisting the two wires of a signal cable. For each wire the electromotive force generated in one twist of the cable is directed opposite to the electromotive force generated in the adjacent twist, as shown for the blue wire. Because of the series connections these electromotive forces result in a net potential difference of  $\approx 0$  V over the total twisted pair. This beneficial effect is maximum when the magnetic field is uniform. Even better is the coaxial cable as shown in the middle as then at all places both a positive and negative loop is present. When capacitive disturbances and ground loops cause problems, the shielded twisted pair is the preferred solution.

result in a total potential difference at the end of the wires of approximately zero Volt, because of the series connection.

Unfortunately often the magnetic fields are not really homogeneous and it is also difficult to keep all twists equal. For that reason the *coaxial cable* is frequently used as shown in the lower drawing of Figure 8.11. In principle the return path of the signal is a tube around the central conductor that carries the signal voltage relative to the grounded tube. A magnetically induced electromotive force at one side of the coaxial cable will be compensated by the induced electromotive force at the opposite side and as long as the resistance in the shield is low this compensation is almost perfect. The word almost is due to the possibility that the magnetic field might not be homogeneous over the cross section of the coaxial cable. A thinner cable is better in that respect but as a side effect the capacitance of the cable would be larger.



**Figure 8.12:** Capacitive disturbances can be reduced by electrostatic shielding. A Faraday shield surrounds the entire sensitive part with conductive material that is either grounded or connected to another trusted constant voltage source. The multitude of parasitic capacitances between the shielding and the instrument will then hardly cause problems, because of the very low relative potential difference.

### 8.3.3.2 Capacitive disturbances

The second source of disturbances is capacitive coupling of electric fields, originating from for instance mains power supply lines and high-frequency circuits of electronic equipment. In principle always some capacitive coupling exists between any electrical wire that carries an alternating voltage and a wire that carries a sensitive signal. Even with small distances and large distance a voltage of 230 V @ 50 Hz creates detectable disturbance values in the mV to  $\mu$ V range in a sensitive, high-impedance electronic circuit as can be observed by touching the input of an audio amplifier by hand.

In principle the capacitive coupling can be approximated as a current source because of the high impedance of the small capacitance value between the source and the sensitive element. This means that the level of the resulting voltages depends on this capacitive coupling with the source impedance of the sensing element and the input impedance of the next element. Often these impedances are both quite high.

Likewise with the magnetic coupling, the capacitive coupling creates voltages in both wires with a common-mode disturbance part that can be cancelled by using a differential amplifier in the next element. Also in this

case the situation for both wires is often different because of the different locations of the wires.

Fortunately it is very well possible to shield the sensitive element by means of a conductive layer, creating a *Faraday shield* around the element that blocks out electric fields as shown in Figure 8.12. This shield can be continuous but also small perforations are often allowed for cooling purposes depending on the disturbance frequency. Very high frequencies can however even penetrate in very small holes, depending on their wavelength!

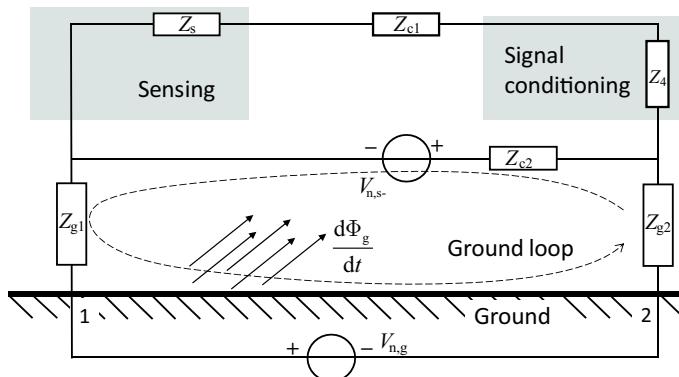
Capacitive coupling can cause a problem with coaxial cables. In principle the signal wire is well shielded by the return wire but this return wire receives all capacitive disturbance signals and the resulting voltage appears at the corresponding input of the next element. A first solution is often to choose that input as the grounded input of the measurement system, short circuiting these disturbance currents to a constant reference voltage. This is mostly sufficient for less critical situations but sometimes the impedance  $Z_g$  of this zero Volt ground point is not low enough, resulting in a small detectable disturbance on the input.

### 8.3.3.3 Ground loops

An even worse situation occurs when the two sides of the connecting cable are both grounded. At first sight one might think that this will short circuit the capacitive disturbances better but this additional connection creates a *ground loop* with a very low impedance, where magnetic disturbances will create a voltage that induces significant currents in the loop impedance. Reducing  $Z_{g,1}$  would even make things worse, because the increased current by the reduced loop impedance will cause an increased voltage over  $Z_{g,2}$ . These negative effects are even present inside an electronic instrument, where the loops are small. Experienced electronic designers are very keen on avoiding ground loops even on printed circuit boards, by grounding all electronics at one location.

Especially when the distances between the elements in a measurement system become large like in a chemical plant, with cables up to hundreds of metres another problem becomes prominent, the difference in potential of the ground at different locations. This potential difference can be caused by leakage currents of large grounded high-power electrical systems, lightning and magnetic fields.

For these reasons double grounding should in principle always be avoided. Safety regulations however often require the housing of mains fed electronic



**Figure 8.13:** Grounding more than one element of a measurement system will create ground loops that introduce common-mode magnetic disturbances and voltages due to other sources. Even though differential amplifiers can reduce the problem, the related voltage levels can be very high, especially with large distances, so ground loops should be avoided.

systems to be connected to the *safety ground*. In those cases the internal electronics are preferably not connected to this grounded casing but often this connection is present for practical reasons. And even without an internal hard wired connection, the capacitance between the electronics and the case creates a capacitive grounding for high frequencies.

To overcome these problems a combination of the coaxial cable and the twisted pairs can be used, the *Shielded Twisted Pair* (STP). The two wires of the twisted pairs conduct the signal including the return path, like with the unshielded version. By very small and well controlled twists, firmly embedded in the insulation, the magnetic disturbances are minimised, while the shielding has no other function than to cancel the electric fields.

In Section 8.5.4 the optical fibre is presented that can be used when the signals are conditioned to a higher signal energy and digitised to a number. The use of an optical fibre prevents any galvanic connection with the related ground loops currents.

## 8.4 Signal conditioning

It is preferred to first add robustness to the electrical signal by placing a signal conditioning element close to the sensing element because of the often very high sensitivity of the sensing element. When possible, the very best solution is created when this element is integrated with the sensor.

As mentioned in the previous section, signal conditioning elements preferably consist of an ideal differential amplifier with an infinite common-mode rejection ratio without offering a load to the sensing element. Furthermore, the signal conditioning element can also contain filters and modulation principles to reduce the impact of interference.

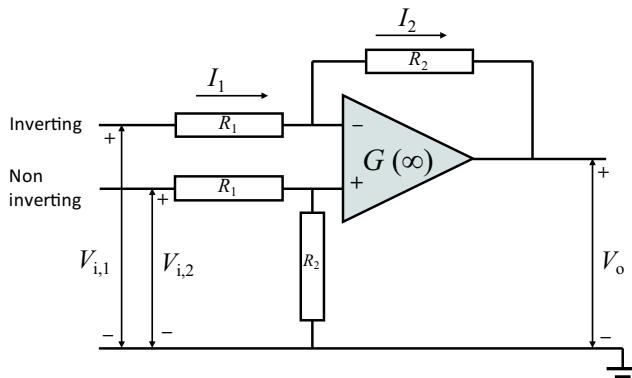
In this section first the instrumentation amplifier will be presented as a true differential amplifier, followed by a presentation of filtering with a focus on modulation techniques.

### 8.4.1 Instrumentation amplifier

Generally the input of a signal conditioning element consists of one or more operational amplifiers. In case of a less critical situation a non-inverting amplifier can be sufficient, when only the impedance of the source is too high for transporting the signal.

When the sensing element acts like a current source as will be shown later with a light sensitive diode, an inverting amplifier with a low input impedance can be applied to convert the signal to a voltage with a low output impedance.

As explained in the previous section however, in most cases a high common-mode disturbance voltage is present in the signal, due to capacitive coupling and ground-loops. In those cases a true differential amplifier is necessary to cancel this common-mode signal. Also in case of a sensor with a Wheatstone bridge a differential amplifier is needed. Unfortunately however the simple differential amplifier that is based on a single operational amplifier as shown in Figure 8.14 has the disadvantage of a difference of loading of the two inputs. The non-inverting input has a simple resistive input impedance, but the current in the inverting input depends both on  $R_1$  and on the voltage at the + terminal of the operational amplifier. This voltage at the + terminal is equal to the non-inverting input voltage and as a result, the current at the inverting input depends on the voltage at the non-inverting input. This would be no problem when both  $V_{i(1)}$  and  $V_{i(2)}$  are delivered by a source with a source impedance that is significantly lower than  $R_1$ . In many sensor



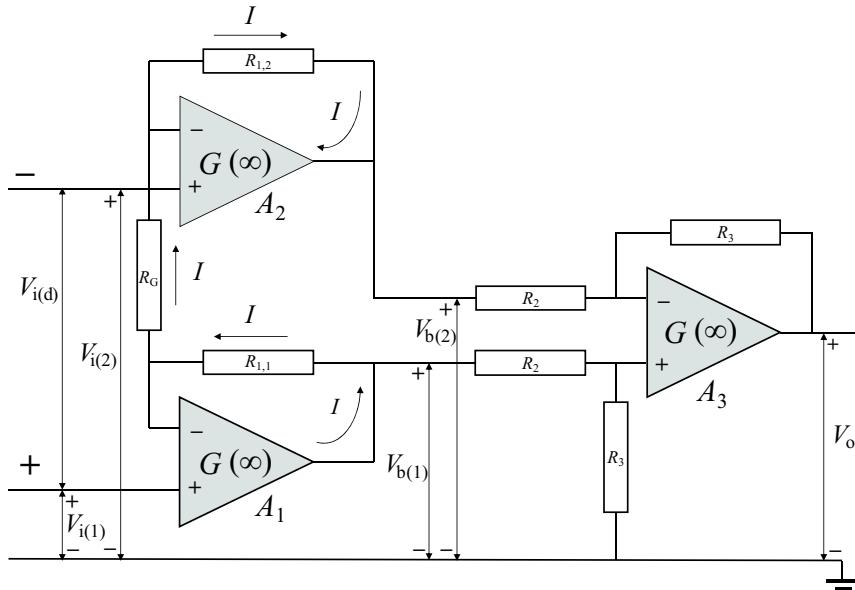
**Figure 8.14:** A simple differential amplifier with one operational amplifier suffers from different loading of the input signals. The current of the non-inverting input is determined by  $R_1 + R_2$ , while the current at the inverting input is determined by  $R_1$  and the voltage at the non-inverting input.

elements this is however not the case and the input impedance has to be increased by adding additional buffer amplifiers before the actual differential amplifier.

Another reason for introducing additional amplifiers is the need to apply only small resistor values, because of noise. The noise in a resistor depends on the resistor value and the temperature as was shown in Section 8.2. Also the input-current noise of an operational amplifier results in a noise voltage at its output that is determined by the feedback resistor value. In practice both inputs of an operational amplifier should “see” a resistive impedance of less than  $1\text{ k}\Omega$  to achieve the minimum noise level and such a low value of the load impedance is not preferred for many sensing elements.

A third reason to adapt the simple differential amplifier configuration is based on the common-mode rejection ratio. In practice it is very expensive to apply resistors with a better tolerance than 0.1 %. As a consequence, with these resistor tolerances, the amplification of the inverting and non-inverting input of the differential amplifier will also differ approximately this same amount. As a result, the common-mode amplification will be only a factor  $10^3$  below the differential-mode amplification, which is equal to only  $-60\text{ dB}$ . This is far worse than the standard CMRR of an operational amplifier, which is often better than  $-100\text{ dB}$ .

Figure 8.15 shows an amplifier that uses two additional operational amplifiers at its inputs to solve these problems. This configuration is called an



**Figure 8.15:** An instrumentation amplifier consists of two non-inverting amplifiers with a standard differential amplifier and one resistor  $R_G$  to set the differential gain. The indicated current direction corresponds with the indicated sign of the input voltage  $V_{i(d)}$ .

*instrumentation amplifier*, a name based on its main application in sensitive measuring instruments.

Instead of simply using two voltage follower buffer amplifiers, the input amplifiers are configured as high-gain non-inverting voltage amplifiers in order to improve the common-mode rejection ratio. The second difference is the resistor  $R_G$  between the negative inputs of the two non-inverting amplifiers that replaces the normally used resistors from the negative input of each operational amplifier to ground.

The instrumentation amplifier has the following working principle.

All operational amplifiers have negative feedback. Under normal conditions, the operational amplifier will do whatever it can do to adapt its output such that the minus input of the matches the plus input. As a consequence the voltage over  $R_G$  will match the differential input voltage  $V_{i(d)}$  of the two non-inverting amplifiers. When the input currents of the operational amplifier are neglected, the outputs of both amplifiers will deliver a current through  $R_G$ ,  $R_{1,1}$  and  $R_{1,2}$ , equal to  $I = V_{i(d)}/R_G$ . When  $R_{1,1} = R_{1,2} = R_1$ , the

intermediate voltages after the non-inverting amplifiers equal:

$$V_{b(1)} = V_{i(1)} + IR_1 = V_{i(1)} + V_{i(d)} \frac{R_1}{R_G} \quad (8.28)$$

and:

$$V_{b(2)} = V_{i(2)} - IR_1 = V_{i(2)} - V_{i(d)} \frac{R_1}{R_G} \quad (8.29)$$

This means that the common-mode voltage of  $V_{b(1)}$  and  $V_{b(2)}$  is equal to the common-mode voltage of  $V_{i(1)}$  and  $V_{i(2)}$  while the differential voltage  $V_{i(d)} = V_{i(1)} - V_{i(2)}$  is amplified with a factor  $R_G/R_1$  for both  $V_{b(1)}$  and  $V_{b(2)}$ . As a consequence the common-mode rejection ratio of this combination is improved with this factor  $R_G/R_1$ . For that reason most of the gain of an instrumentation amplifier should be realised in these non-inverting input amplifiers, while the resistors of the differential amplifier part are often chosen to be equal resulting in a unity gain of the second part. With this equal resistor setting the output voltage becomes equal to:

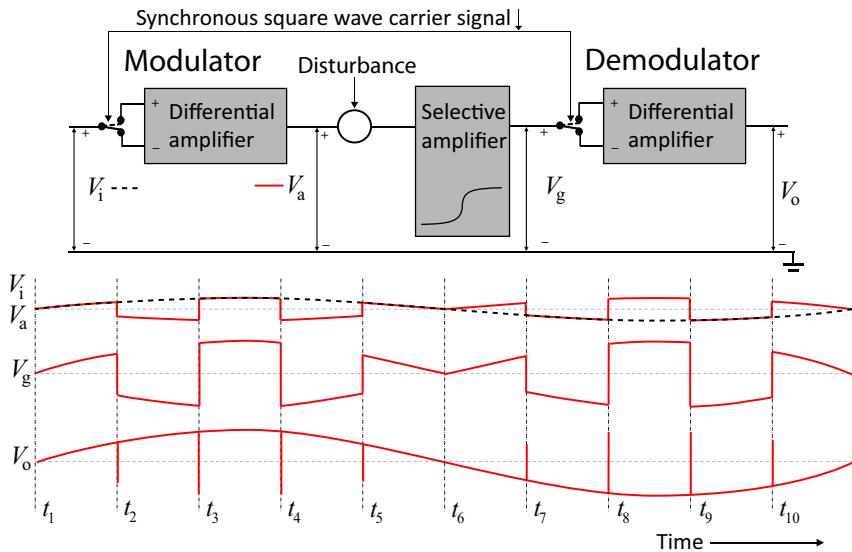
$$V_o = V_{b(1)} - V_{b(2)} = V_{i(1)} - V_{i(2)} + 2V_{i(d)} \frac{R_1}{R_G} = V_{i(d)} \left( 1 + 2 \frac{R_1}{R_G} \right) \quad (8.30)$$

Fully integrated versions of such instrumentation amplifiers exist, where only the external gain-setting resistor  $R_G$  must be added to realise a fully robust signal conditioning amplifier.

### 8.4.2 Filtering and modulation

Even with a careful design it is not always possible to reduce all sources of random errors to an sufficiently low level. In that case it is preferred that these errors are first reduced by filtering, before the signal is further processed in the measurement system. Simple filtering can be very useful if the disturbance signal has a different frequency spectrum than the useful signal. In that case the active filter configurations can be used like described in Section 6.2.7 of Chapter 6. Examples are the reduction of the high-frequency part of wide-band noise of electronic components and single frequency disturbances can be filtered by a selective band reject filter. The second-order “taming” low-pass filter in the PID-controller that was presented in Chapter 4 also serves this same purpose.

Unfortunately often the disturbance signal occurs in the same frequency band as the measurement signal. This is especially the case with low-frequency excess noise because most mechanical measurements have also



**Figure 8.16:** Synchronous modulation and demodulation is achieved by alternately switching the signal input  $V_i$  between the inverting and non-inverting input of a differential amplifier. The resulting high-frequency signal  $V_a$  is amplified by a selective amplifier, resulting in  $V_g$ . A demodulator, switching synchronous with the input switch, inverts the amplified signal again into a low-frequency signal  $V_o$ . The disturbance is cancelled in the selective amplifier.

a strong low-frequent component in their spectrum. This can be solved by combining the measurement signal with a high-frequency *carrier* signal and only amplify the resulting high-frequency combination. This combining action is called *modulation* and several types of modulation are applied. For measurement systems, *amplitude modulation* (AM) is frequently used. It is well-known from radio transmission and will be explained more in detail. Other principles include frequency modulation that is also used in radio transmission, pulse-width modulation that was introduced in Chapter 6 for power amplifiers and phase modulation that will be presented in Section 8.8 with incremental optical measurement systems.

#### 8.4.2.1 AM with square wave carrier

One principle of amplitude modulation is shown in Figure 8.16. The high-frequency carrier signal is a square waveform, alternately switching the input signal between the inverting and non-inverting input of a differen-

tial amplifier. The output of the modulator is a high-frequency amplitude modulated square wave signal and can be filtered below the carrier wave frequency without losing information. This filtering is done by means of a *selective amplifier* with a high-pass frequency transfer function. As a result, disturbances at the connection between the differential amplifier and the selective amplifier will not arrive at the input of the demodulator. The demodulator consists of a similar combination of a switch and a differential amplifier as the modulator, restoring the original sign of the signal input by switching synchronous with the modulator. This special demodulation method is called *synchronous demodulation* because of the synchronous switching of the modulator and demodulator.

Because of the sharp transients of the modulated square wave signal, the amplifier has to be able to preserve the very high frequencies of the square wave spectrum. For that reason, the use of this version of amplitude modulation is restricted to relatively low-frequency measurement signals.

For higher frequencies a sine wave carrier frequency is preferred. Instead of switching, the modulation with a sine wave carrier frequency is achieved by multiplication. In fact also the switching modulator acts like a multiplier of the momentary magnitude of the signal input with the momentary magnitude of a square wave signal.

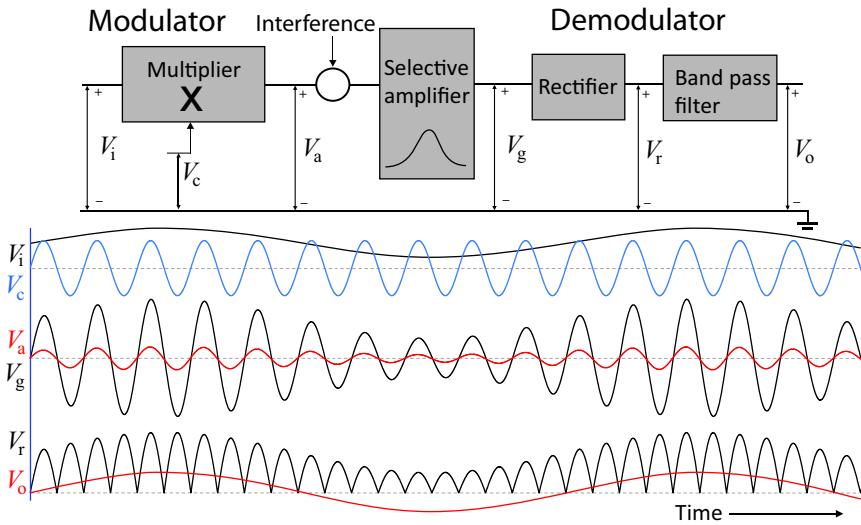
#### 8.4.2.2 AM with sinusoidal carrier

In radio transmission the sine wave carrier signal is multiplied with a combination of a DC bias voltage and the music signal, as shown in Figure 8.17. This method has the advantage that demodulation is fairly easy and for that reason it is also useful in measurement systems, when the measurement signal does not contain a DC value.

As a first step, the measurement signal is added to the DC bias voltage, large enough to keep the voltage always unidirectional. By multiplying this combined DC+AC signal voltage with the carrier signal, the resulting high-frequency signal will never get a zero amplitude. This modulated signal can be amplified with a selective amplifier with a band-pass transfer function around the carrier frequency.

Demodulation is achieved by simple rectification, followed by a second band-pass filter around the measurement signal frequency. This filter cancels the DC bias voltage and the high-frequency remains of the carrier frequency.

To determine the transfer function of the selective amplifier it is useful to look at the simple mathematics around this modulation principle. When the



**Figure 8.17:** Amplitude modulation by multiplying the input signal with the carrier signal. The input signal consists of the measurement signal and a DC bias voltage. As a result, the amplitude of the multiplier output will never be zero. After amplifying and filtering, the demodulation is done with a simple rectifier and a band-pass filter.

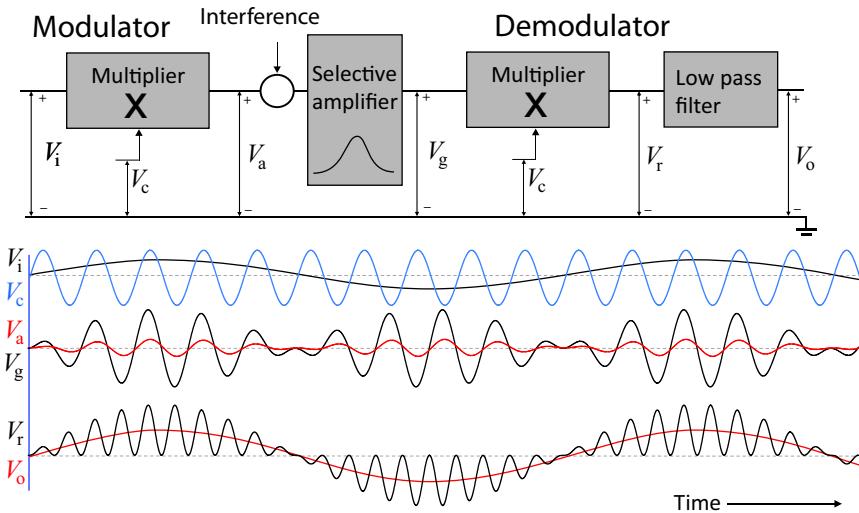
input signal with the DC bias voltage equals  $V_i = \hat{V}_i(1 + \sin \omega_i t) = V_i(1 + \sin 2\pi f_i t)$  and the carrier signal equals  $V_c = \hat{V}_c \sin(\omega_c t) = \hat{V}_c \sin(2\pi f_c t)$  the multiplication results in:

$$\begin{aligned} V_a &= V_c V_i = \hat{V}_c \hat{V}_i \sin(2\pi f_c t)(1 + \sin(2\pi f_i t)) \\ &= \hat{V}_c \hat{V}_i \left( \sin(2\pi f_c t) + \frac{\cos(2\pi(f_c - f_i)t) - \cos((f_c + f_i)t)}{2} \right) \end{aligned} \quad (8.31)$$

This means that the resulting signal contains three frequencies, the carrier signal frequency  $f_c$  and two frequencies equally spaced at both sides of the carrier frequency at a frequency difference equal to  $f_i$ . These frequencies are called the *side-bands* around the sampling frequency and they both contain the information of the measurement signal at  $f_i$ .

With this result it can be reasoned that an input signal that contains different frequencies up to  $f_{\max}$  requires the bandwidth of the selective amplifier to range from  $f_c - f_{\max}$  to  $f_c + f_{\max}$ . In principle one of the two side-bands could be filtered out and restored with special measures afterwards, but that procedure is too specialistic for the purpose of this book.

A drawback of this simple scheme with rectifier demodulation is the frequency limitation to AC signals. The synchronous demodulation of the first



**Figure 8.18:** Synchronous modulation and demodulation with a sinusoidal carrier signal enables the application of an input signal without a DC bias voltage. The sign of the amplified measurement signal is retrieved by a second multiplication in the demodulator with the same carrier frequency.

example with a square wave signal did not have that drawback. Fortunately it is also possible to apply synchronous demodulation with a sine wave carrier signal, as shown in Figure 8.18. It results in the amplification of frequencies from 0 Hz and above.

The input signal is not added to a DC bias voltage, because also any DC content of the input signal should be maintained. The resulting signal after modulation has a zero amplitude at  $V_i = 0$  and shows a phase reversal, when  $V_i < 0$ . Demodulation by rectifying would result in a frequency doubling with a DC offset, so the demodulation is done by multiplication with the same carrier signal.

The average of the resulting signal is equal in shape as the input signal and can be recovered by a low-pass filter. This method is also called *phase selective detection* or *homodyne detection* and will return later in this chapter with different sensing principles.

## 8.5 Signal processing

After the signal conditioning element, a strong and robust signal is present that can be used in an analogue control system. Modern control electronics work predominantly in the digital domain, which makes it necessary to convert the momentary values of the electrical signal into a sequence of data that sufficiently represents the measured value.

Digital systems are characterised by the fact that they work with very strictly timed events, synchronised by means of an accurate clock frequency generator, mostly based on a quartz crystal oscillator.

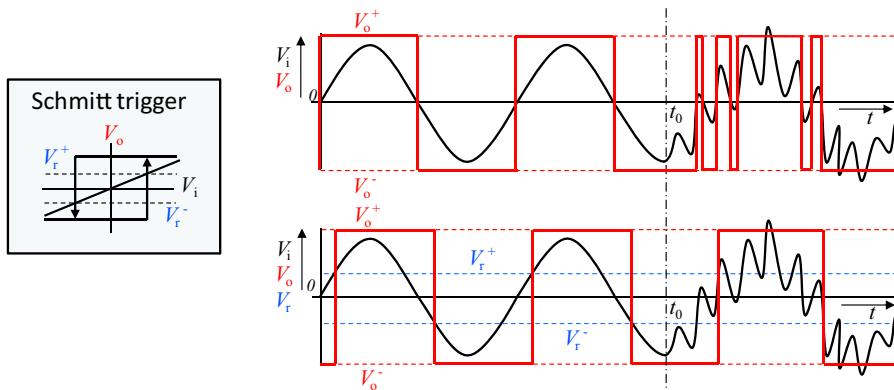
When only the frequency of a periodic signal is important, the analogue signals can be directly converted into the digital domain. This *frequency-to-digital conversion* process is achieved by means of a comparator, a one bit *bistable* element that outputs either a “zero” or a “one” depending on the voltage level at the input. In chapter 6, a special version of such a comparator, the Schmitt trigger, was used in the pulse-width modulator of a switched mode power amplifier. In measurement systems the Schmitt trigger is used to suppress disturbances.

With real analogue-to-digital conversion the magnitude of a value needs to be digitised and that is done by taking a momentary value of the measurement signal and generate a number to the magnitude of that sample. In the following sections, first the Schmitt trigger will be presented, followed by the process of creating sampled numerical data from the magnitude of a measurement signal.

### 8.5.1 Schmitt trigger

A comparator is a differential amplifier with an infinite gain. The output is either a fixed positive value when the plus-input is at a higher voltage than the minus-input or a fixed negative value in the other case. A Schmitt trigger is a comparator with a well defined hysteresis level.

Figure 8.19 shows the functionality of a Schmitt trigger with different signals. When first examining the upper graph, without any hysteresis, the output voltage  $V_o$  changes its value any time the input signal crosses the 0 V reference level. This configuration without hysteresis is only useable, when the input signal would be without any noise, which is never reality. The effect of noise is shown after  $t_0$ , where a noise signal is added. The noise will cause additional output transitions at the zero crossing with a higher frequency than the input signal. When examining the lower graph,



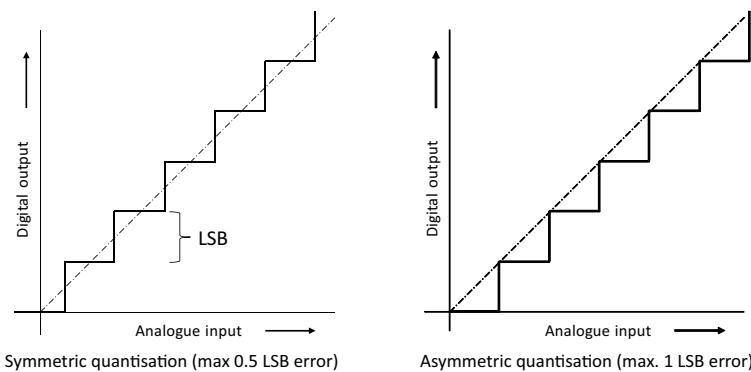
**Figure 8.19:** A “Schmitt trigger” is a comparator with a defined level of hysteresis that converts a periodic signal into a square wave with a reduced sensitivity for noise. The upper right graph shows the different voltages when no hysteresis is present, while the lower right graph shows the voltages with a hysteresis level of  $V_r$ . The input signal is sinusoidal and after  $t_0$  noise is added to show the difference. Note the phase shift due to the hysteresis.

the influence of the hysteresis is clearly observed. Starting at a situation where  $V_o$  is negative, the input signal rises until it surpasses the value of  $V_r^+$ . At that moment  $V_o$  will transit to the positive side. Only when the input signal drops below the now negative value of  $V_r^-$ ,  $V_o$  will transit to the negative side again. This hysteresis has two effects. Firstly an additional phase shift occurs between  $V_o$  and  $V_i$  depending on the amplitude of  $V_i$  and the hysteresis level but the secondly effect is more important because the frequency of  $V_o$  will not be influenced by the noise as long as the peak to peak value of the noise remains below the hysteresis level.

Comparators in both inverting and non-inverting configurations are frequently applied and even when not specifically mentioned, they mostly are provided with some kind of hysteresis. Only when the hysteresis level is clearly defined the comparator deserves its name as Schmitt trigger.

### 8.5.2 Analogue-to-Digital conversion

Digital values are represented in a binary format, because computers work with binary logic circuits. The elementary digits of a binary number are called *bits*, with only two values, zero and one. A binary number consists of a summation of the value of  $n$  bits, with  $n$  being an integer  $> 0$ . In the



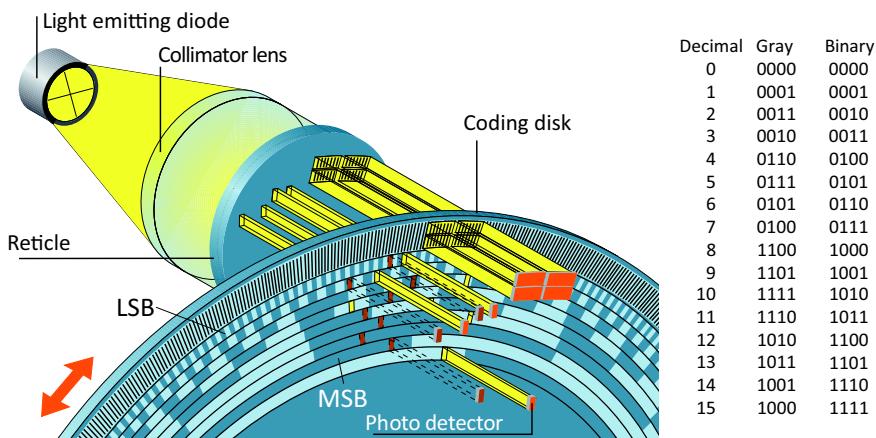
**Figure 8.20:** The incremental nature of digitisation introduces a quantisation error that can range between maximum 0.5 to 1 times the least significant bit depending on the method of quantisation.

mathematical sense, the value of bit  $k$ , where  $0 \leq k < n$ , is equal to  $2^k$ . The bit with  $k = 0$  is called the *Least Significant Bit* (LSB) and the bit with  $k = n - 1$  represents the *Most Significant Bit* (MSB).

When digitising measurement values, the magnitude represented by the most significant bit is equal to half the total range of the measurement value. The magnitude of the least significant bit represents the smallest increment that can be distinguished between different measurement values. This incremental property of digitisation determines the first source of errors in a measurement system, the *quantisation error* as shown in Figure 8.20. Depending on the quantisation method, the error can amount up to a maximum of one least significant bit (LSB).

### 8.5.2.1 Gray code

Although digital numbers offer in principle an unambiguous representation of a value, in practice its application can cause problems. This is due to the fact that with several incremental steps the change of one increment is represented by a change of more than one bit. This effect is most significant in a situation, where the value is changed from just below half the range to just above half the range or the other way around. With for instance an eight bit number, increasing with one increment this would mean a change from 01111111 to 10000000. If for some reason not all bits switch at the same moment, this can lead to erroneous data. A well-known configuration, where this situation can occur, is the direct digital measurement of an angular

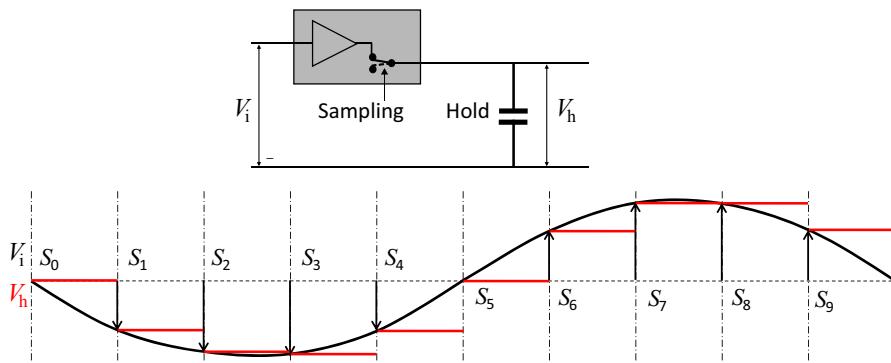


**Figure 8.21:** With direct measurement of digital data, a binary code can cause problems by not synchronised changing of all related bits. This is solved by using a Gray code that switches only one bit at any increment, like in this example of an absolute rotary optical encoder. At the right a 4-bit coding is shown as example.

(Courtesy of Heidenhain)

or linear position by means of a multitude of light beams through holes in a coding disk that represent the different bits. Figure 8.21 shows an example of such a rotary measurement system. Instead of a binary code this measurement system uses a *Gray code*, named after the American physicist Frank Gray. The Gray code is designed such that only one bit is switched at any incremental change. As shown in the example at the right of the figure the least significant bit switches at all uneven decimal numbers, so once in every two increments. The next bit switches once in every four increments starting from zero to one at the second decimal value, the third bit switches on every eight increments, starting at the fourth decimal value and so on. The use of a Gray code prevents erroneous data due to the tolerances in the alignment of the slots on the coding disk with the photo sensitive detectors. The example of the rotating disk shows the effect of a position alignment error but also errors can occur, when not all bits change at the same time due to a different switching speed of the bits. This timing problem can be solved by a *latch*, a memory that holds a digital value for a certain time and can be read out at the moment that all data is stable. This is applicable in systems that work with a well defined clock frequency, where all digital data are synchronised.

Most converters that will be presented in the following sections, need several



**Figure 8.22:** Sampling a signal is done with a switch, while a capacitor will hold the value constant until a new sample is taken. A buffer amplifier before the sampling switch prevents excessive loading of the input signal by the hold capacitor.

clock cycles to determine all bits of the digital value of an analogue sample. During this process each determined bit is stored in the latch and only when all bits are found, this number is transferred to the digital processor in the last clock cycle.

### 8.5.2.2 Sampling of analogue values

The incremental nature of digital signals is both related to the value and to the time. Numbers can not change gradual and are only valid over a certain period. This fact implies that digital numbers represent *samples* of a momentary value of the analogue signal. Depending on the method of analogue to digital conversion, the sampling frequency can be constant, synchronous with the clock frequency of the digital controller or variable depending on the signal value. The latter can be synchronised afterwards by means of a memory where the data are written asynchronous but the readout is synchronised.

Samples are by definition only exact at the moment of sampling and this automatically means that intermediate information is lost. In case of a varying analogue signal, it seems logical that the sampling frequency should at least be larger than the largest frequency in the signal.

It is important to see, what the limitation really is.

### 8.5.2.3 Nyquist-Shannon theorem

The low limit of the sampling frequency in a perfect analogue to digital conversion is determined by the following key rule of any conversion process:

Any conversion is only perfect, when it can be reversed to reproduce the original input signal.

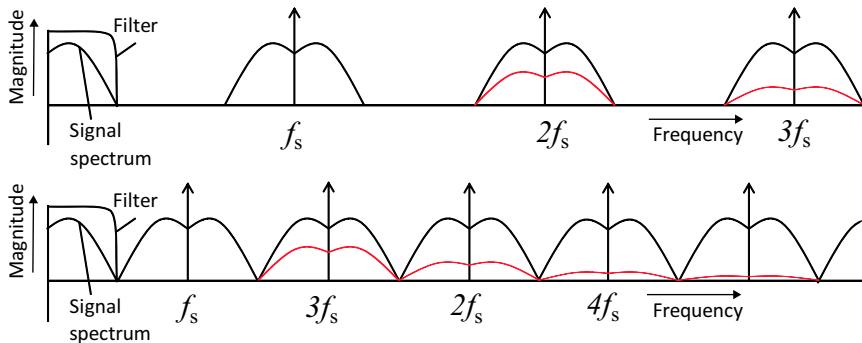
With this rule in mind, first the sampling process itself is considered. Figure 8.22 shows how sampling can be done in practice. The samples of the signal are taken with a switch that shortly connects a capacitor to the signal voltage at the output of a buffer amplifier. During this sampling moment the capacitor is charged to the momentary signal voltage. After opening the switch, the sampled voltage remains stored in the capacitor until the next sample is taken. As will be shown with the A to D converters further in this chapter, this *sample-and-hold* functionality is necessary for those converters that need the input signal to be constant during the conversion process. Other converters do not have that requirement as they average the input signal over each sample.

One observation that can be made from the figure is the time delay that is caused by the “hold” functionality. In principle this delay corresponds with the fact that information is lost between the samples. The average time delay by sampling is half the sampling period and unfortunately this delay can never be restored! This in itself is an important determining factor for the minimum sampling frequency when the measurement is used to control a process.

Even more important is the effect of sampling on the frequency spectrum. In principle, without the hold function, sampling alone is equal to amplitude modulation of a repeated impulse signal, while the holding action results in a kind of square wave modulation. With reference to the previous section on signal conditioning with amplitude modulation, the frequency spectrum of a sampled signal also shows “side bands” around the sampling frequency. With a sampled impulse signal this pattern is repeated at n-times the sampling frequency due to the wide frequency spectrum of an impulse.

This phenomenon is illustrated in Figure 8.23, where also the effect of the “hold” functionality is shown with a low-pass filter character, reducing the magnitude of the spectrum at higher harmonics of the sample frequency.

The upper graph in the figure shows the frequency spectrum in case the sampling frequency is much higher than the maximum signal frequency. The original signal spectrum could be simply recovered by filtering away all



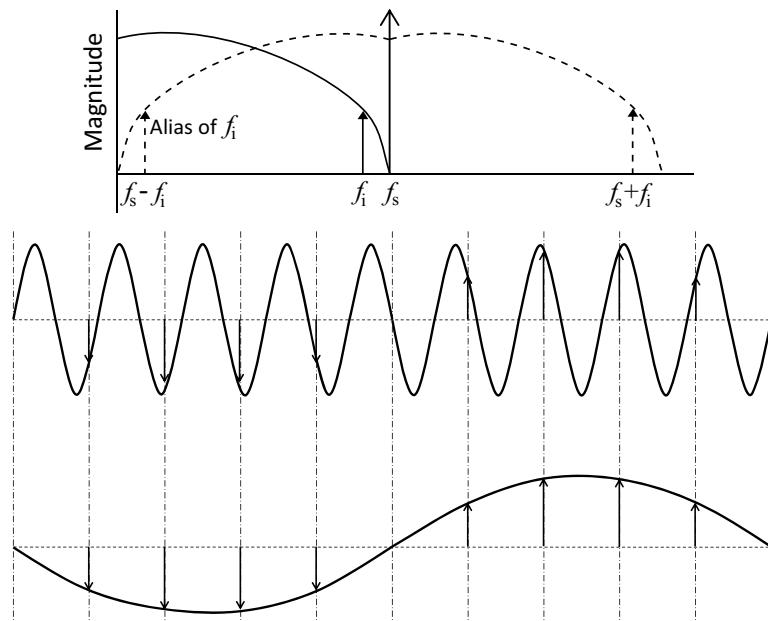
**Figure 8.23:** Sampling a signal will create mirroring side bands of the frequency spectrum of the signal around  $n$ -times the sample frequency, where  $n$  is an integer  $> 0$ . When  $f_s$  is smaller than two times the maximum signal frequency, overlap between the spectra will occur, impairing the possibility to reproduce the original signal. The effect of the “hold” action at higher frequencies is shown in red.

sampling related signals above the maximum signal frequency.

The lower graph shows the effect when the sampling frequency is exactly two times the maximum signal frequency. The lowest frequency of the first sideband around the sampling frequency is equal to the maximum signal frequency and only an infinitely sharp filter could prevent that nearby higher frequencies from the side band are introduced in the original signal. With an even lower sample frequency this can no longer be avoided and it becomes not possible anymore to recover the original signal.

## Aliasing

Figure 8.24 shows what happens when the sampling frequency is for instance equal to the maximum frequency of the input signal spectrum. The side bands around the sampling frequency overlap with the frequency spectrum of the analogue input signal and any frequency of this spectrum will get a corresponding alias frequency that is lower than the original frequency, equal to the frequency difference  $f_d = f_s - f_i$ . When using these samples for reconstruction two frequencies would result, the original and its alias, which has given the name to this effect. In case the sample frequency is higher than two times the signal frequency this difference frequency remains larger than the maximum signal frequency and when all frequencies above the original signal frequency are filtered out, only the original frequency would remain.



**Figure 8.24:** An alias frequency is generated when sampling with a frequency  $f_s$  that is lower than twice the maximum signal frequency. The frequency component  $f_i$  of the analogue input spectrum combines with the sampling frequency into a much lower frequency  $f_s - f_i$ , the alias frequency. This aliasing effect is also demonstrated in the lower two graphs, where the same samples correspond with two different analogue signal frequencies.

Based on the occurrence of aliasing, the Nyquist criterion for sampling is postulated by the same Harry Nyquist, who created the Nyquist plot. This criterion was later improved by the American mathematician Claude Elwood Shannon (1916 – 2001) into the Nyquist-Shannon sampling theorem by incorporating the effects of noise. The theorem defines the *Nyquist frequency*  $f_N$  as half the sampling frequency and the theorem reads as follows:

When sampling, the signal frequency spectrum is not allowed to contain frequencies above the Nyquist frequency.

This theorem corresponds with the following expressions:

$$f_{i,\max} < f_N = \frac{f_s}{2} \iff f_s > 2f_{i,\max} \quad (8.32)$$

### 8.5.2.4 Filtering to prevent aliasing

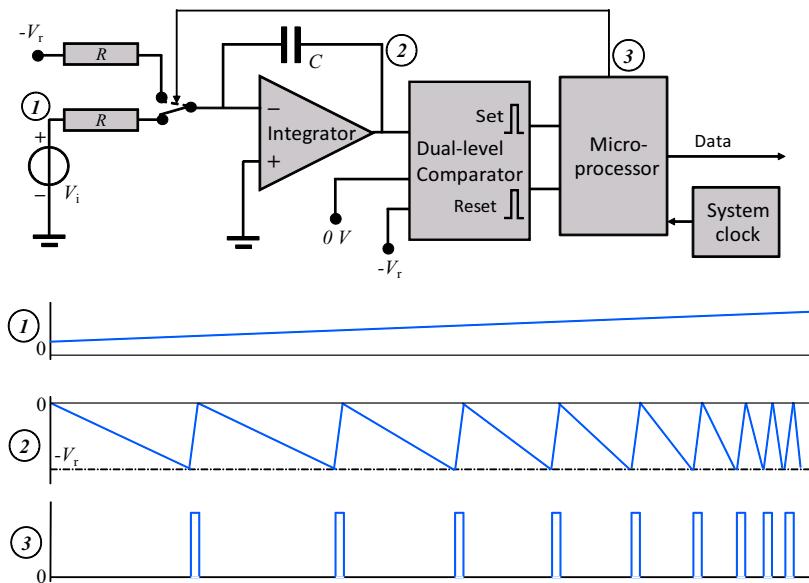
In most cases, signals in a measurement system consist of a large spectrum of frequencies. When  $f_{i,\max}$  is the maximum frequency of interest in this spectrum, the sampling frequency  $f_s$  must at least comply to the Nyquist-Shannon theorem and remain above  $f_N$ .

In practical systems it is however better to take some more margin especially in feedback controlled motion systems, where the phase margin determines the dynamic performance. This is caused by the need to apply filters twice in the loop. The first filter, the anti aliasing filter, is used to prevent that non-relevant frequencies higher than  $f_N$  enter the sampling process. This filter can never be infinitely steep so it is better to keep the sampling frequency a little bit above the  $2f_{i,\max}$ . The other filter is used when the analogue signal needs to be reconstructed, before the analogue amplifier part after the controller. The higher frequency components in the sampled signal must be filtered away as otherwise these frequencies would cause unwanted effects in the amplifier and actuator. All low-pass filters show a phase delay that increases with their steepness, especially when overshoot is not allowed (high damping). For these reasons it is preferred to choose a higher sampling frequency with a lower-order filter.

A nice example of the application of the Nyquist-Shannon theorem is the digitisation of audio signals for a Compact Disc. The sample frequency of the first CD-players and recordings was set at 44 kHz, only with a small margin of 10 % above two times the maximum audible frequency of 20 kHz. This caused a lot of criticism at the introduction of the CD, because it resulted in a “harsh” sound quality induced by the sharp higher-order filters needed to prevent sound artifacts by aliasing. In present days, the increased capability of digital electronics enabled even in a simple CD-player a digital over-sampling algorithm by higher-order interpolation to a multiple of the original sample frequency, like 96 or even 192 kHz. When reconstructing the original signal this high sampling frequency allows the application of simple second-order Bessel filters with a constant phase delay, which is far more pleasant to listen to.

## 8.5.3 Analogue-to-digital converters

In this section three different *analogue-to-digital converters* (ADC) will be presented. First the most simple version is shown, the *dual-slope ADC* that generates sampled values of a variable unidirectional input voltage, asynchronous with the heart beat of the further digital processing. It is used



**Figure 8.25:** Dual-slope analogue-to-digital converter. A microprocessor measures the time in which the output voltage of an integrator changes from 0 V to a negative reference voltage level  $V_r$ . This time is proportional to the positive input voltage and is calibrated with the integrator time to 0V caused by switching the input of the integrator to the negative reference voltage  $V_r$ .

in less critical situations like with controllers for central heating of buildings. This converter does not require a separate sampling and hold function and can be realised with only few components around a microprocessor at a very low cost. The second converter is the frequently applied *successive-approximation* AD converter. It combines a high precision with an average speed and has long been a standard in A-to-D conversion. This converter is not as fast as the third converter, because it requires a separate sampling and hold function and the quantisation error is maximum 1 LSB. The third converter is the *Delta-Sigma converter*. Its principle is derived from the dual-slope ADC, but adapted such that it creates synchronous samples of alternating voltages at a very high sampling frequency.

### 8.5.3.1 Dual-slope ADC

The working principle of a dual-slope ADC is shown in Figure 8.25.

This converter is only suitable for signals with a single polarity. A switch at the input of an inverting integrator connects the input of this integrator to either the positive analogue input signal or to a fixed negative reference signal with a known value. A dual-level comparator is used to detect the precise moment that the output voltage of the integrator passes either below a negative reference voltage  $V_r$  or above a positive voltage of 0 V. As soon as one of these events occurs, the comparator will give a set- or a reset-pulse to a microprocessor. A set-pulse corresponds to the moment of reaching 0 V and after that event the microprocessor will command the switch at the input of the integrator to connect to the input voltage of the AD converter. The output voltage of the integrator will then fall at a proportional speed of the analogue input voltage and the microprocessor measures the time for the integrator to reach the reference value, which triggers a reset-pulse from the comparator. As soon as this reset pulse is received, the microprocessor will command the switch at the input of the integrator to connect to the reference voltage and starts measuring the corresponding time for the integrator to rise to 0 V again.

The ratio between the two obtained time slots is equal to the ratio between the absolute values of the reference voltage and the average input voltage during the integration. With the known value of the reference voltage this ratio can be converted by the microprocessor into a voltage value.

This dual-slope converter is the most simple of all configurations as it can be made with only a few electronic parts. It is often used in microprocessor-controlled devices, like the controller of a central heating system, and frequently the integrator is replaced by an RC-network while the inputs of the microprocessor are often used as the comparator inputs. In that case the microprocessor is programmed to compensate the resulting non-linearity and other anomalies that are related to this low-cost solution.

The resolution of this converter is limited by the amount of clock pulses of the microprocessor in one integrator cycle, which reduces its use to less time-critical applications. If for instance the clock-frequency of the microprocessor is 1 MHz, a 1 ms integrator-cycle would contain only  $10^3$  pulses. Also with an input voltage close to 0V the cycle time increases, which makes this converter less suitable for feedback controlled motion systems that need a low and constant phase delay.

### 8.5.3.2 Successive-approximation ADC

A more precise and relatively fast conversion procedure is achieved with the successive-approximation analogue-to-digital converter. With this method, the output of a *digital-to-analogue converter* (DAC) is compared with the input signal, while an efficient algorithm is used to get the right digital number in the least possible steps. Figure 8.26 shows this working principle with a DAC based on a *R-2R ladder-network*.

The *R-2R* ladder-network is best understood when starting at the first section with the least significant bit (LSB) switch  $S_1$ . The source impedance at point (1) is two times  $2R$  in parallel, which results in a Thevenin impedance of  $R_{Th} = R$ . This is indifferent of the setting of switch  $S_1$ . With both settings a low impedance path to ground is created at the switch, as a voltage source has a zero source impedance.

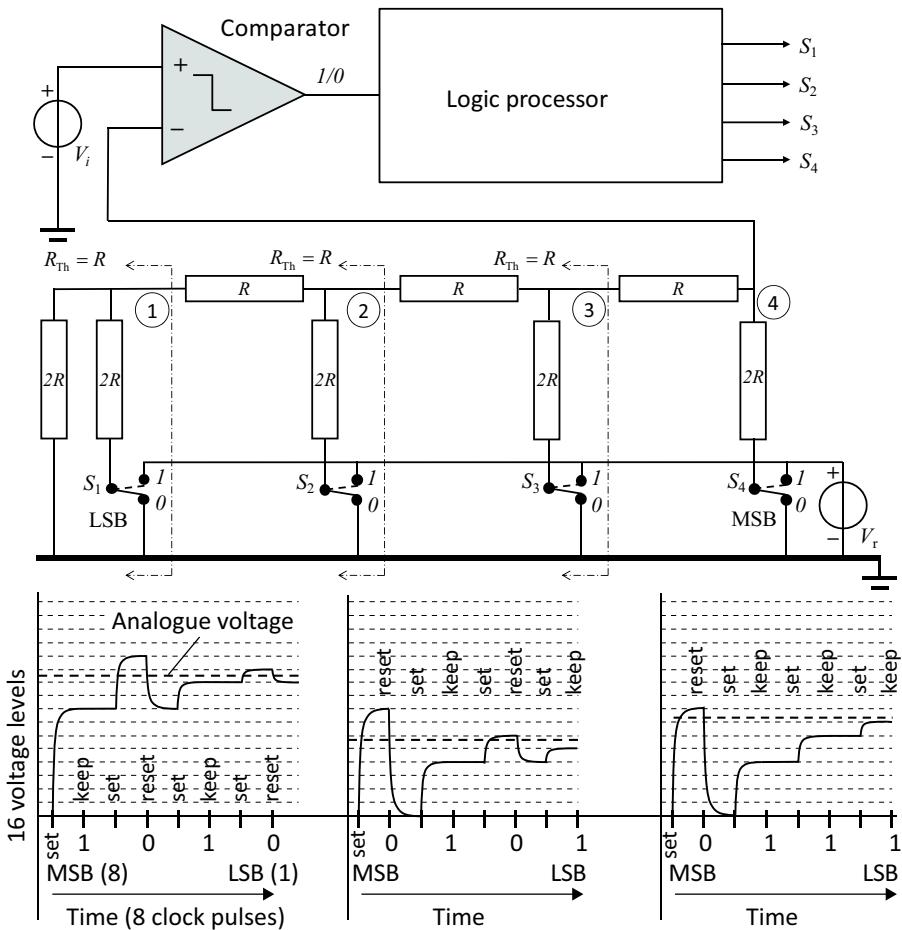
The Thevenin voltage  $V_{th(1)}$  of this first section at point (1) equals either zero or  $0.5V_r$ , depending on the setting of  $S_1$ . The Thevenin impedance of the first section is loaded by the second section and creates a voltage divider with the resistances  $R$  and  $2R$  of the second section, where the  $2R$  resistor is connected to ground at  $S_2$ , either directly or via the voltage source. This voltage divider results in an attenuation of the Thevenin voltage of the first section with a factor two, so either  $0\text{ V}$  or  $0.25V_r$ .

The equivalent Thevenin source impedance of this first and second section combined also equals  $R$  as two  $2R$  resistors are connected in parallel to ground like with the first section. When switching  $S_2$  to  $V_r$  the resulting Thevenin voltage at point (2) will be the sum of the voltage from  $S_1$  ( $0$  or  $0.25 \times V_r$ ) and the voltage from  $S_2$  ( $0$  or  $0.5 \times V_r$ ) resulting in a total voltage of  $0$ ,  $0.25$ ,  $0.5$  or  $0.75 \times V_r$ , depending of the settings of  $S_1$  and  $S_2$ .

This reasoning can be repeated for the next stages to get the total voltage at point (3) and point (4). While in the figure only 4 bits are shown, an accurate multi-bit DA-converter can be designed with only two values of resistors, which is very suitable for realisation in a precision integrated circuit.

The logic circuit for the AD-converter works as follows:

- First the most significant bit (MSB) is switched on with an output voltage of the DAC of half the full scale. When this results in a higher output voltage than the analogue input voltage, the output of the comparator will go from 1 to 0 and the processor decides to reset the MSB back to 0. Otherwise the processor decides to keep the MSB to 1.
- Then the second significant bit is switched on and the same decision process takes place: “Keep bit setting if output of DAC is smaller than



**Figure 8.26:** A 4-bit Analogue-to-Digital converter, based on the principle of successive approximation, with an R-2R DA converter, a comparator and a logic processor. The lower graph shows in three examples how the four bits are sequentially determined in eight set and decide (keep or reset) steps.

input, reject when larger”.

- This step is continued for every bit in sequence of lower significance until the least significant bit is defined.

The main advantage of the successive-approximation ADC is its absolute resolution and inherent linearity, which is mainly determined by the applied DAC. As long as the DAC has enough time to stabilise after each step, the additional errors can remain below the value of the least significant bit. Note

that this stabilisation of the voltage at the DAC is always necessary because of any parasitic capacitances that are always present in any electronic circuit. This is indicated in the figure by the gradual slope after the (re-)setting of any bit. This timing requirement inherently leads to a the drawback of this method as it requires many steps for a high number of bits. During these steps the input voltage has to remain constant until the entire procedure is finished. This means, as was stated before, that after the moment of sampling the momentary value of the input voltage has to be “remembered” as long as necessary with a sample and hold circuit. Also this process needs time as the conversion process can only start, when the hold value is stabilised.

The consequence of this reasoning is that very high speed electronics are needed, which in practice always implies an increase in cost.

### 8.5.3.3 Delta-Sigma ADC

The last example of an ADC, the Delta-Sigma ( $\Delta\Sigma$ ) analogue-to-digital converter, is based on another principle that combines the cost benefits of the dual slope converter with a much higher speed and makes use of extreme oversampling. Oversampling means that the signal is sampled with a frequency that is significantly higher than necessary according to the “Nyquist” criterion of two times the maximum signal frequency. This allows the use of rather simple electronics, no sample and hold and only a one bit DAC with just two values,  $V_r^-$  and  $V_r^+$ . Figure 8.27 shows the basic circuit diagram of this converter and the signals with a sinusoidal input voltage.

The core of a  $\Delta\Sigma$ -ADC is of a  $\Delta\Sigma$ -modulator that converts the input voltage in a “bitstream” consisting in a timed stream of 1 bit coded signals. A digital processor, the “decimator” converts this bitstream in a stream of multi-bit decimal data. The  $\Delta\Sigma$ -modulator is a variation on the pulse-width modulator that was described in Section 6.3.3 for a switched-mode power amplifier. The block diagram of Figure 8.27 shows the main components of a  $\Delta\Sigma$ -modulator:

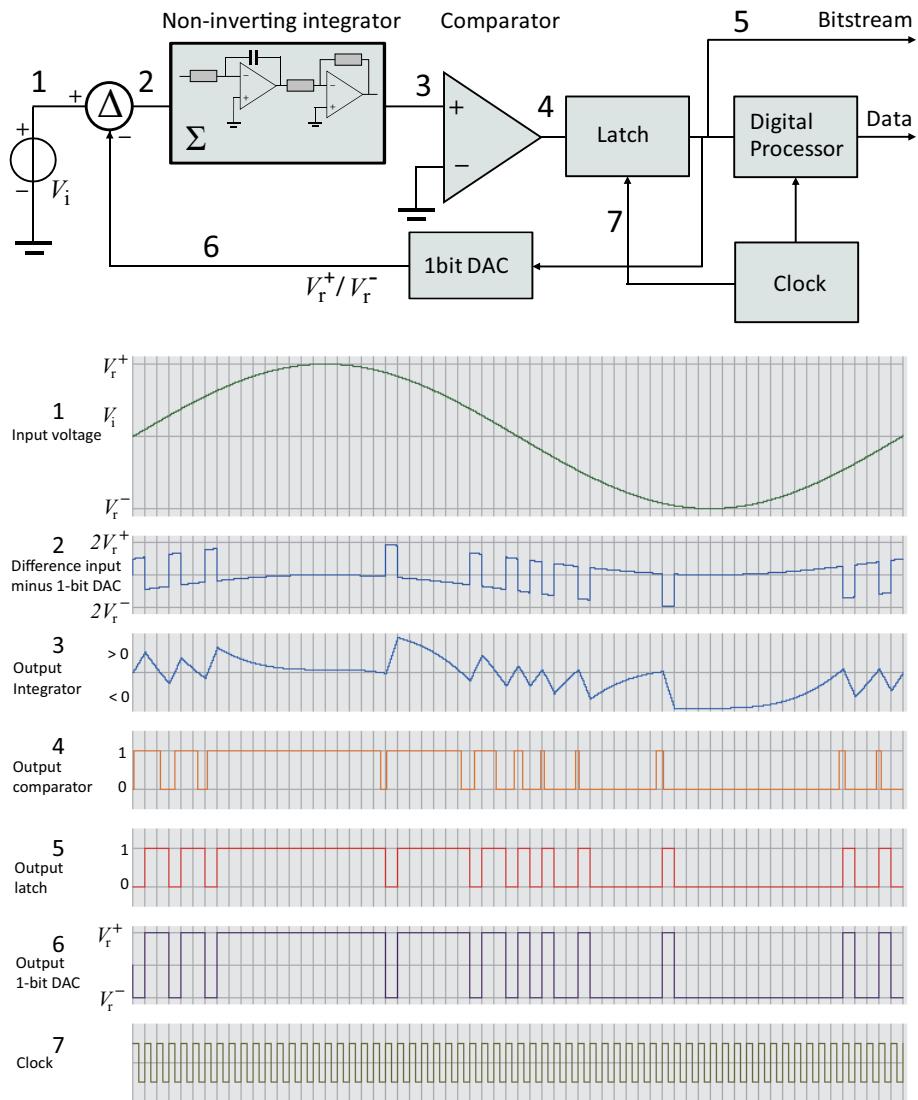
- A differential amplifier, shown as a circle, that gives the difference between the analogue input voltage and the output of the 1-bit DAC.
- A non-inverting integrator that consists of an inverting integrator and an inverting amplifier with a gain of one.
- A non-inverting comparator with a very small hysteresis.
- A latch that transfers the momentary value at its input tot the output at the positive edge of the clock.
- A digital processor that creates digital data from the output of the latch.
- A clock that generates the timing sequence of the conversion process.
- A one bit AD-converter with two output states.

The non-inverting integrator integrates ( $\Sigma$ ) the difference ( $\Delta$ ) between the momentary value of the input signal and the output of the 1-bit DA-converter. The output of the non-inverting integrator will rise when this difference is positive or fall when this difference is negative. Like with the pulse-width modulator, the comparator is used to detect the moment that the output of the integrator crosses a reference value, but in this case the hysteresis

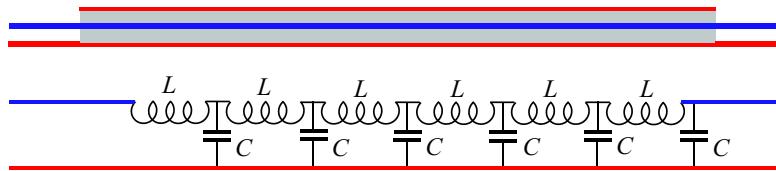
is very small and the reference voltage is 0 V. Another difference is that the moment of switching does not immediately take place after a change in the output of the comparator but the decision moments are based on the clock with a latch that transfers the output of the comparator to the digital processor at a positive edge of the clock. The output of the latch will set the 1-bit DAC to give  $V_r^+$  when the output of the comparator gives a 1, which corresponds to a positive voltage at the output of the integrator. The 1-bit DAC will be set at  $V_r^-$  when the comparator gives a 0 at a negative output voltage of the integrator. The resulting output of the latch is a so called *bitstream*, a continuous stream of bits at a very high frequency. This bitstream can be converted in normal digital numbers by a suitable decimation algorithm giving multi-bit decimal data at fixed intervals.

In principle the loop with the integrator-inverter-comparator-latch-DAC-integrator acts like a negative feedback loop with three times an inversion. As a result of the integration, the steady state error is zero and the average value of the output of the 1-bit DAC will be equal to the inverted value of the input voltage. With a little bit of imagination this is visible in trace (6) in Figure 8.27. As a consequence, the reconstruction of the original signal is possible by simple low-pass filtering of this bitstream. This is comparable with the output filtering in a switched mode PWM power amplifier.

This direct reconstruction principle is applied in the ultimate audio format, “Super Audio CD” with “Direct Stream Digital” registering maximum 22 kHz audio with 64 times oversampling which gives  $64 \cdot 44 \cdot 10^3 = 2.8224$  MHz clock sampling. It achieves a dynamic range of 120 dB in the audible frequency range, while it is allowed to register frequencies up till 100 kHz at a reduced dynamic range without any aliasing effects because of the high sampling frequency.



**Figure 8.27:** Delta-Sigma analogue to digital converter. A non-inverting integrator integrates the difference between the input signal and the output signal of a one bit DA converter. A positive output of the integrator results in a one and a negative output in a zero at any clock cycle. The resulting “stream” of bits has an average analogue value equal to the analogue input signal.  
 (Courtesy of Uwe Beis)



**Figure 8.28:** A cable with two wires, like a coaxial cable, is equivalent to a series of small inductors and capacitors. This combination represents a characteristic resistive impedance, equal to the required resistive load to get  $Q = 1$ .

### 8.5.4 Connecting the less sensitive elements

Even though the signal conditioning element serves to create a robust signal, it is necessary to transport this signal with care in a coaxial or twisted pair cable. Also the digital data after the A to D conversion should be transported with a careful choice of the connecting cables. Next to the reason to prevent interference, especially with very long distances, another reason for a well defined connection are the high-frequency properties of a connecting cable.

#### 8.5.4.1 Characteristic impedance

Any combination of two wires will have a certain total capacitance and inductance value. When examining a connecting cable with two wires more in detail, such a cable can be modelled as an infinite series of infinitesimal LC-filter elements. Such a combination behaves like a *transmission-line* or *wave-guide*, an electrical equivalent to the travelling wave in the rope as was presented in Chapter 2. Just like with the mechanical example, a signal at the input of the cable will arrive with some delay at the output, but what is more important is that the signal will be reflected back towards the input, where it will again be reflected. This can cause standing waves and resonances in the signal.

It was shown in Chapter 6 that an LC-filter can be damped by means of a resistor and in the same way the reflection of a connecting wire can be prevented when it is terminated by means of a resistor with a value that would result in a quality factor  $Q = 1$ . With Equation (6.40) the value of this

*characteristic impedance* can be calculated:

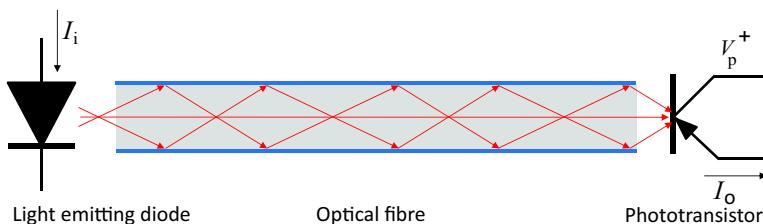
$$Q = R_c \sqrt{\frac{C}{L}} = 1 \quad \Rightarrow \quad R_c = \sqrt{\frac{L}{C}} \quad (8.33)$$

This value is equal for any length of the same cable as both  $L$  and  $C$  depend proportional to this length. It can be mathematically proven that a cable with infinite length represents an input impedance that is equal to this characteristic impedance. By terminating a non-infinite cable with its characteristic impedance, the cable will behave dynamically like it is infinitely long. This means that an optimal connection is achieved, when both the output impedance of the preceding element and the input impedance of the succeeding element are chosen equal to the characteristic impedance of the cable.

When that condition is not met, a short cable might possibly not raise much problems as it will only give an undamped response at a very high natural frequency. For example a coaxial cable with a characteristic impedance of  $50 \Omega$  has a capacitance of approximately  $50 \text{ pF}$  per metre. Based on these values the inductance over one metre would be equal to  $L = R_c^2 C = 125 \text{ nH}$ . When all capacitance is approximated to be concentrated at the end of the cable and all inductance is in series, this combination would act as an LC filter with a natural frequency of  $\omega_0 = \sqrt{1/LC} = 4 \cdot 10^8$ , which is approximately  $60 \text{ MHz}$ . No reflections will occur as long as the signals that are transported do not contain any frequencies at or above  $\omega_0$ . Actual wired data connections, however, already transmit frequencies above  $1 \text{ Gbit}$  per second so in that case even one metre of cable should be terminated with the right impedance in order to prevent reflections and high-frequency attenuation. Longer cables will increase the problem by lowering the natural frequency proportional to the length of the cable.

The full derivation of the related equations is beyond the scope of this book but some observations are interesting enough to be added here.

At first sight the equal output and input impedance will result in a factor two attenuation of the signal. This however is not bad at all as first this level corresponds with the maximum power that the preceding element can deliver, as was shown in Chapter 6. This is optimal from a disturbance point of view as that is related to the interfering energy relative to the energy in the useful signal. Furthermore, because of the transmission line character, even frequencies above the original natural frequency will be transmitted equally well because at any position in the cable the signal “sees” only the resistive characteristic impedance in two directions and the resistance is not frequency dependent.



**Figure 8.29:** Transfer of a digital signal with a fibre coupling prevents electric disturbances and errors by ground loops. It consists of a light emitting diode that transmits light through the optical fibre towards a phototransistor.

#### 8.5.4.2 Non-galvanic connection

Although well defined wired connections are very suitable for many applications it still suffers from the problems with ground loops and other disturbances, that were previously presented. A real solution of all these problems is based on the transfer of information by means of light. In Section 6.3.3.6 of Chapter 6 the opto-coupler was introduced, consisting of a light emitting diode (LED) and a light sensitive phototransistor that are electrically insulated from each other. A current in the LED will give a proportional light output that creates a proportional current in the transistor. The current amplification ratio of the transistor results in a current level in the same order of magnitude as the current in the LED. Also light sensitive diodes can be used. They have a higher switching speed but also a reduced current because they fail the current amplification of the phototransistor. A clear advantage of the opto-coupler is the galvanic insulation that completely cancels any kind of ground loops, although at very high frequencies some capacitive coupling still remains. Because of the significant temperature dependent non-linearity, the application of opto-couplers is mainly limited to the digital domain, so after the signal processing element. A special version of the opto-coupler is the fibre coupling of Figure 8.29. The insertion of an optical fibre between the LED and the phototransistor completely cancels the capacitive coupling but more important is the possibility to transport the signal over very long distances without any electromagnetic interference. When necessary even a strong semiconductor laser can be used as the light source. This technology has become the de-facto standard for data transfer over very long distances like with internet.

## 8.6 Short-range motion sensors

The first category of position and motion sensors are those sensors that are used to determine the relative position of two objects at short ranges and these sensors are often called a *proximity detector*. Typical applications include the boundary detection for long-range actuated systems. Also the sensors for fixating a position by means of a servo system like in a magnetic bearing belong to this category. Often only the accuracy at the zero position is relevant while the sensitivity over the total measurement range is relevant for the gain of the servo-loop.

The second category consists of those sensors that register motion like velocity and acceleration. They are applied in many different fields like in geophysics to detect earthquakes and in precision equipment for vibration control.

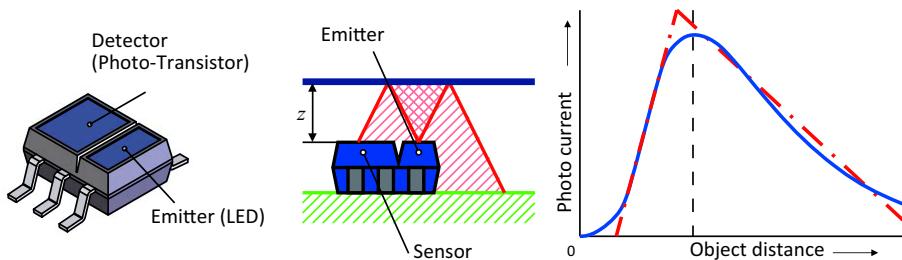
Because of this wealth of possibilities only a few examples are selected as these are representative for the entire field

### 8.6.1 Optical sensors

The first example of a proximity detector is shown in Figure 8.30. It consists of a light emitting diode and a photo-transistor sensor in one housing “looking” in the same direction. The working principle of this optical proximity detector is based on the difference in irradiance levels detected at the photo-transistor as a function of a distance of the object. It is interesting to see that there are two areas that can be used with an approximately linear relationship between the position and the current. This is due to the fact that at zero distance the object will obscure the LED sensor combination and at infinite distance no light is reflected anymore. Somewhere in between there is a maximum of captured light at the detector with two slopes at each side. The slope at the near side has the highest sensitivity.<sup>5</sup>

In order to achieve some level of accuracy both modifying and interfering error sources of this system must be determined. First of all the current level of the LED has influence as it is directly proportional to the amount of emitted light. The overall shape of the curve of Figure 8.30 will be unaffected but the magnitude is linearly proportional to the amount of light from the

<sup>5</sup>If this detector is applied in a servo system it is necessary to be aware that the two slopes work with an opposite sign. This means that a stable system working on one slope will be unstable due to phase sign reversal at the other slope. This necessitates special precautions, when the system is out of range and the stable point has to be retrieved.



**Figure 8.30:** An optical proximity detector consisting of a Light Emitting Diode with a sensor that determines the irradiance of the reflected light from an object. The relation between the current of the sensor and the distance of the object (blue line) shows two opposite approximately linear slopes (dashed red line) that are both applicable for position sensing.

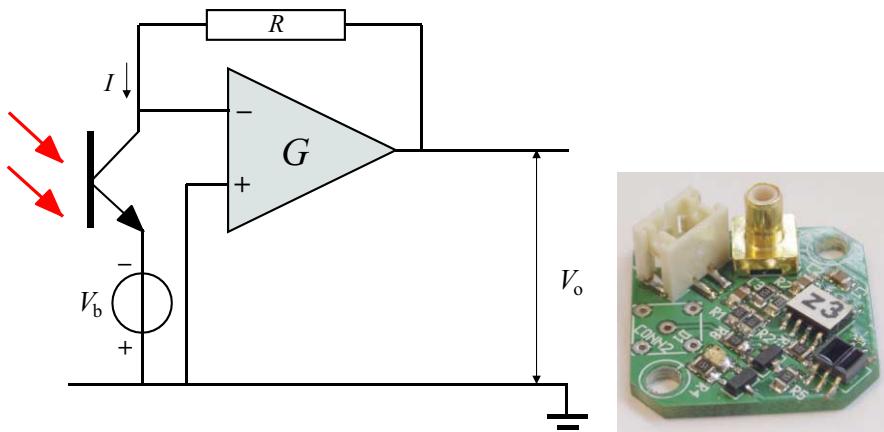
(Courtesy of Jasper Wesselingh)

LED. This means that the current in the LED is both a modifying input as it changes the slope (sensitivity) and an interfering input as it changes the value at any position. The effect of this can be reduced by comparing the current with a reference current that is proportional to the current in the LED. Also supplying the LED with a constant current source as presented in Chapter 6 is a suitable method to reduce the influence of the LED.

Other error sources include the reflectivity of the object, the angle of the object, noise in the detector and ambient light that is detected by the phototransistor. The influence of the ambient light can be reduced by using an optical filter in order to only detect the wavelength of the LED light. To limit the noise of the detector an amplifier with a low input noise level must be located as close as possible near the sensor. Figure 8.31 shows an example of such a circuit which is also optimised for high-frequency behaviour as the transistor is directly connected to the virtual ground at the minus input of the amplifier. This means that the current from the transistor is transformed into a voltage at the output of the amplifier without causing a voltage change over the transistor itself, thus avoiding high-frequency current from the collector to the base by the internal parasitic Miller capacitor.

### 8.6.1.1 Position sensitive detectors

Instead of a single photo-transistor or a photo-diode also a sensor consisting of a multitude of sensing elements can be used, giving both information on the irradiance and on the position of the light spot that is directed to the



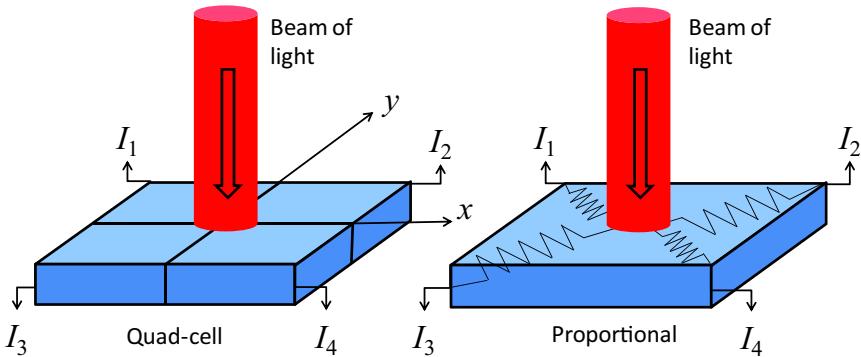
**Figure 8.31:** A photo-transistor works optimal with reduced current noise, when it is supplied with a constant bias voltage  $V_b$ . The virtual ground at the minus input of the operational amplifier cancels the AC voltage over the transistor which gives a better high-frequency behaviour.  
(Courtesy of Jasper Wesselingh)

device. An example of this principle is the *four-quadrant detector* with 4 photo-diodes, also called a *quad-cell*, as shown at the left side of Figure 8.32. Also an analogue version of this four quadrant detector exists as shown on the right side. The total area of this detector is sensitive to the incident light but the resulting current is transported to the four external contacts by means of a material with a high resistivity. This causes the current of each contact point to be inverse proportional to the distance to the incident spot of the light.

With both types of PSDs the ratio between the current of each output is ideally related to the position of a light spot that is exposed on these sensors according to the following equations:

$$\begin{aligned} x &= \frac{(I_2 + I_4) - (I_1 + I_3)}{I_1 + I_2 + I_3 + I_4} \\ y &= \frac{(I_1 + I_2) - (I_3 + I_4)}{I_1 + I_2 + I_3 + I_4} \end{aligned} \quad (8.34)$$

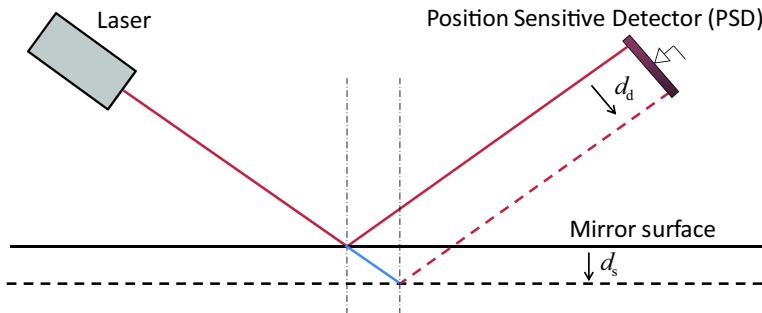
By dividing the difference of the currents by the sum of all four currents the resulting value is independent of the irradiance of the light beam which would otherwise be a modifying input. In reality, the relation between the position and the found values also depends on the size of the light beam, the distribution of the light inside the beam and the character of the sensor.



**Figure 8.32:** The four quadrant position sensitive detector either consists of 4 separate photo diodes as shown at the left side or of one large photosensitive area with high resistive paths to the output wires as shown at the right.

Generally a four-quadrant detector is used as a zero-sensor in a feedback system like in the Optical pick-up unit of a CD-player, as around zero the errors due to diameter and light distribution are minimal. A proportional sensor is less sensitive for these deviations but it shows more noise due to the high resistivity.

A more elaborate example of such a sensor is the CCD sensor as used in digital cameras and the Shack-Hartmann wavefront sensor from Chapter 7. In order to avoid a large number of wires a CCD sensor has a dedicated electronic circuit, the charge coupled device, that makes it possible to read out a row of photo-diodes in series by sequentially transferring the accumulated charge of each photo-diode (the integral of the current over the sampling time) to its neighbour like buckets of water that are transferred by people in a row to empty a well.

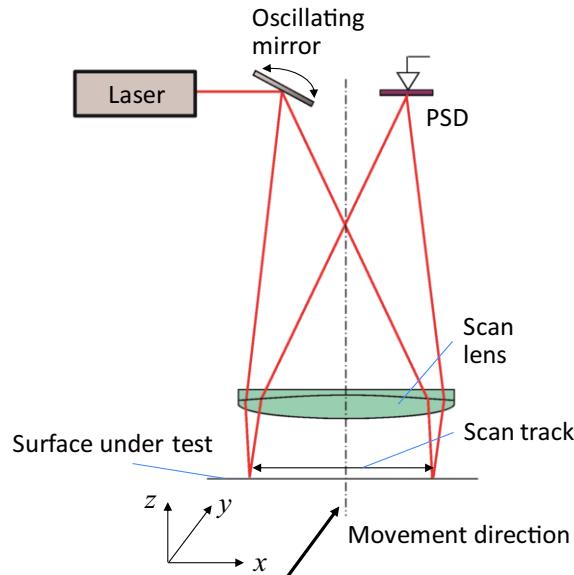


**Figure 8.33:** Surface displacement measurement with a position sensitive detector by means of triangulation. A shift of the surface ( $d_s$ ) will cause a shift ( $d_d$ ) of the laser beam on the PSD.

### 8.6.1.2 Optical deflectometer

A position sensitive detector can be used to measure several mechanical properties. By deflecting a beam of light off a surface and projecting it on the PSD it is possible to determine the angle and/ or the position of this object as shown in Figure 8.33. This method is called *triangulation* because the real displacement depends on the angles and can be calculated by trigonometry. In this simple set-up the angle can not be distinguished from the displacement of the surface. This can be solved by focusing the beam on the surface and imaging that focal spot by a lens on the PSD. In that case the angle will have no influence anymore.

An example where imaging is used to detect the angle of a surface instead of the height is the *Deflectometer* of Figure 8.34. This instrument is used to measure the surface shape of a reflective surface. It works as follows: A laser beam is directed to a mirror that oscillates around the point of incidence of the laser beam. Due to this oscillation, the laser beam will be reflected in different directions between the outer positions as shown in the picture causing a scanning light spot on the scan lens. This lens is positioned such that the rotating point of the mirror is in its focal plane which means that the rays from the laser beam after the lens are always running parallel. The angles of these rays are only determined by the position of the oscillating mirror relative to the optical axis so they are independent of the momentary angle of the mirror. The surface that must be measured reflects these parallel rays back to the lens where they will be focused again to create an image of the original spot on the rotating mirror. Because the rotating mirror is positioned just left of the optical axis

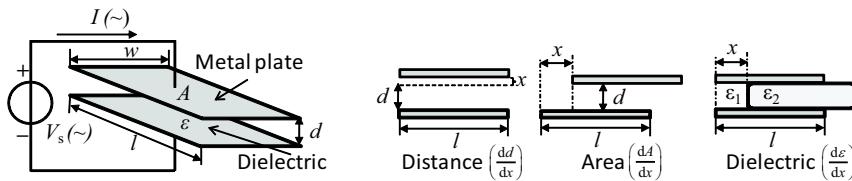


**Figure 8.34:** Optical deflectometer that measures the shape of a surface by a scanning beam and suitable optics that images the spot of the oscillating mirror to the PSD via the surface. Local angular variations of the surface will result in position differences of the spot on the PSD.

the image will be positioned at the same distance<sup>6</sup> right of the optical axis, where the PSD sensor is located. Because the spot of the rotating mirror is imaged on the PSD, the position of this image is not influenced by the angle of the mirror as long as the surface that must be measured is perfectly flat. When that is not the case the parallel incident rays on the surface will not be reflected parallel anymore and an angular deviation at a certain location on the surface will result in a change of position of the spot on the PSD. When the corresponding electrical signal is synchronised with the oscillation of the mirror, this position signal of the PSD gives information about the slopes at different locations on the surface and by calculation the shape of that surface can be derived.

---

<sup>6</sup>This is a 1:1 optical system as both image and object are in the focal plane and a positive lens inverts the sign of the position relative to the optical axis.



**Figure 8.35:** Three types of capacitive position sensing by changing the position of two plates of a capacitor in two directions or changing the dielectric properties as function of the displacement.

## 8.6.2 Capacitive position sensors

In Section 6.1.2.2 of Chapter 6 it was shown that a capacitor consists of two parallel plates, separated by an electrically non-conductive dielectric layer with a capacitance of:

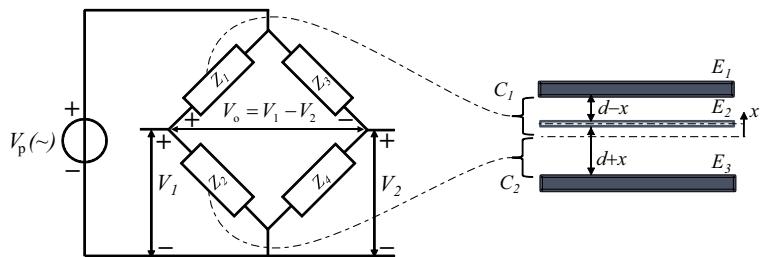
$$C = \frac{\epsilon A}{d} \quad (8.35)$$

With  $\epsilon = \epsilon_0 \epsilon_r$  and

$$\epsilon_0 \approx 8.8541878176 \cdot 10^{-12} \quad (8.36)$$

For air  $\epsilon_r$  equals approximately one but as will be shown further in this chapter, this assumption is a cause for errors.

This inverse relationship of the capacitance with the distance is the reason that this physical property is used to determine displacements between two objects that are positioned close together. An AC source is necessary to determine the capacitive value with the complex version of Ohm's law ( $V = Z \cdot I$ ). In Figure 8.35 it is shown that not only the mutual distance between the plates can be used to measure displacements but also the other two terms in the simple relation can be made variable corresponding with a displacement. In case the displacement of one of the plates is in its own plane the distance remains the same but in this case the shared surface is changed while in the third example a separate part with another dielectric value is inserted between the plates which also gives a change in the capacitance according to the position of this part relative to the plates. Each solution has its own advantages. A variation of the distance gives generally a higher sensitivity than the parallel movement of the plates but the latter has a better linearity which can be concluded from the respective



**Figure 8.36:** A differential capacitive sensor has a linear relation between the displacement  $x$  and the voltage  $V_0$ .

transfer functions of these three examples:

$$1: \text{Variable-distance:} \quad C = \frac{\epsilon A}{d+x} \quad (8.37)$$

$$2: \text{Variable-area:} \quad C = \frac{\epsilon}{d}(A - wx) \quad (8.38)$$

$$3: \text{Variable-dielectric:} \quad C = \frac{\epsilon_0 w}{d} (\epsilon_2 l - (\epsilon_2 - \epsilon_1)x) \quad (8.39)$$

From these three different sensing principles it can be concluded that a measurement according to one principle is also influenced by the other effects. Especially the distance has a very large influence on the sensor in case of the variable-surface and variable-dielectric sensor. With these principles the distance acts as a modifying error source. In case of the variable distance sensor the lateral displacement has less influence especially if one of the electrodes is larger than the other. Because of its high sensitivity, the variable-distance measurement capacitive sensor is preferred, but the non-linearity still is a factor that preferably is avoided. This can be achieved by means of a differential sensor and the previously presented Wheatstone bridge as shown in Figure 8.36.

### 8.6.2.1 Linearising by differential measurement

The differential capacitive sensor consists of two stationary electrodes \$E\_1, E\_3\$ with a moving electrode \$E\_2\$ in between. The capacitance \$C\_1\$ is the combination of \$E\_1\$ with \$E\_2\$ and the capacitance \$C\_2\$ is the combination of \$E\_2\$ with \$E\_3\$. When the surface of the electrodes is equal to \$A\$ the capacitance becomes:

$$C_1 = \frac{A\epsilon}{d-x}, \quad C_2 = \frac{A\epsilon}{d+x} \quad (8.40)$$

The voltage  $V_o = V_1 - V_2$ . When applying the rules of a voltage divider this voltage equals:

$$V_o = V_p \left( \frac{Z_2}{Z_1 + Z_2} - \frac{Z_4}{Z_3 + Z_4} \right) \quad (8.41)$$

In order to get a positive output voltage with a movement in the positive  $x$ -direction  $C_1$  can be located at  $Z_1$  and  $C_2$  at  $Z_2$  to give an increase of  $V_1$ . Another option would be when  $C_1$  is located at  $Z_4$  and  $C_2$  at  $Z_3$  as in that case a positive movement in the  $x$ -direction gives a decrease of  $V_2$ . In this example  $Z_1$  and  $Z_2$  are chosen and  $Z_3 = Z_4$  are two equal resistors to balance the bridge.

With this choice the following equation gives the voltage as function of the displacement:

$$V_o = V_p \left( \frac{\frac{1}{j\omega C_2}}{\frac{1}{j\omega C_1} + \frac{1}{j\omega C_2}} - \frac{1}{2} \right) = V_p \left( \frac{d+x}{d-x+d+x} - \frac{1}{2} \right) = x \frac{V_p}{2d} \quad (8.42)$$

This is a linear and frequency independent relation.

### 8.6.2.2 Accuracy limits and improvements

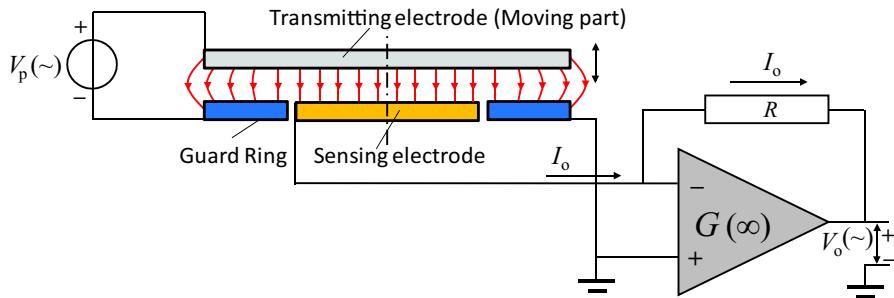
From the previous part it can be concluded that the errors in capacitive measurement sensors are caused by many external influences, summarising:

- The amplitude and frequency of the source signal.
- The noise of the sensing electronics.
- Crosstalk of movements in other directions.
- Thermal effects due to change in size of the electrodes.
- Changes in the dielectric.
- Inhomogeneity in the electric field.
- External electric fields.

It is clear that the electric source has a large impact if the values are not exactly known. Fortunately these electrical parameters can be controlled to high levels of precision and the measurement result can in most case be compensated. This is less the case with the error by the noise of the sensing

electronics. In practice a capacitive sensor has a capacitance in the range of a few pico-Farad which requires a very high excitation frequency to create a sufficiently high current level for an acceptable signal to noise ratio. This sensitivity can be illustrated by using the data of one of the best operational amplifiers on the market regarding noise, the AD797 from Analog Devices. It shows a voltage noise of  $0.9 \text{ nV}/\sqrt{\text{Hz}}$  and a current noise of  $10 \text{ pA}/\sqrt{\text{Hz}}$ . With a 1 kHz bandwidth this is  $\approx 30 \text{ nV}$  and  $\approx 0.3 \text{ nA}$  of noise. When this amplifier is used with a capacitive sensor of 6 mm diameter active area giving a surface area  $A \approx 30 \text{ mm}^2$  and a distance  $d$  of 300  $\mu\text{m}$ , this results in a capacitance of  $\approx 1 \text{ pF}$ . A voltage source of 10V @ 1 MHz will create a current of  $\approx 0.1 \text{ mA}$ , a factor  $3 \cdot 10^5$  above the 0.3 nA noise of the amplifier. With a distance of 300  $\mu\text{m}$  this noise level corresponds with around 1 nm error. When a lower error value is needed it is only possible to decrease the working distance and hence limit the range.

The factors on thermal effects and crosstalk of other movements are directly related to the mechanical construction of the sensor in combination with its surrounding parts. An isothermal measurement condition is required to avoid the thermal errors but in principle it is preferred to design the mechanics around the sensor in such a way that the so called *thermal-centre* is located inside the sensor. The thermal centre in a construction is defined as the place where no local movement is observed at a change of temperature. When the capacitive sensor is located at the thermal centre of the measurement system, a temperature rise does not affect the distance of the plates. This is not always possible to achieve as it requires a thermally symmetrical construction around the sensor where all other parts move away from the thermal centre at a temperature rise. Sometimes the use of materials with a different thermal expansion coefficient can help to achieve this goal. Aside from a linear movement by temperature rotations and other movements between the plates have to be avoided by a stable construction. Changes in the dielectric are present in case the sensor is not used in vacuum which is almost always the case. The relative permittivity of air is a function of its composition and humidity and pressure variations can be observed in the measurement value. Although this relation is expected to be quite straightforward, the humidity level has a more direct effect that is not yet completely understood. It is based on the occurrence of thin layers of water on any object that is in contact with air that contains water. The layer thickness depends on the partial pressure of the water vapour (level of saturation). This layer has a far higher permittivity than air and can in principle be approximated as a continuation of the electrode. This means that the air gap is decreased with the same amount as the water-



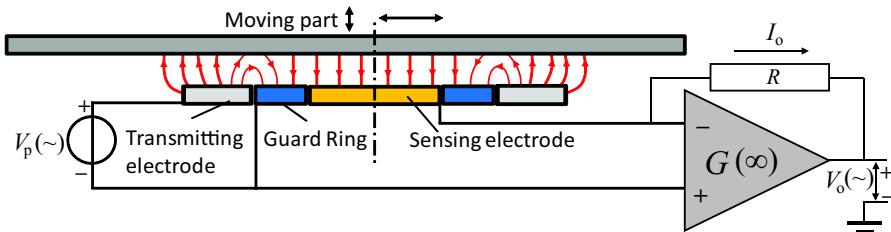
**Figure 8.37:** By using a guard ring the electric field at the sensing electrode is homogenised. The capacitance is measured by virtually grounding the sensing electrode and measuring the current with a transimpedance amplifier.

layer thickness. Much research is spent on this subject with measured thickness's ranging from 1 nm to almost a  $\mu\text{m}$  but no concluding results are found to sufficiently model this phenomenon so that it can be compensated. Though this phenomenon is influenced by the hydrophilic properties of the electrodes, the effect occurs even with a PTFE layer. For extreme precision measurements this means that these have to be done in dry air or vacuum. Short term measurements are less critical as a change in the water layer by the environmental humidity is generally rather slow, in the range of minutes, depending on the geometry and diffusion speed.

Inhomogeneity in the electric field is caused by edge effects as the field lines are bent at the edges of the sensor. This can be reduced by a special configuration as shown in Figure 8.37 for a single ended sensor. It should be noted that this sensor does not work with a Wheatstone bridge but in this case the current of the sensing electrode is measured when grounding this electrode at the virtual ground of a transimpedance amplifier.

At the upper side a large electrode is supplied by the electrical supply with a high-frequency AC voltage relative to ground. The opposing electrode, the sensing electrode is much smaller and is surrounded by a guard ring that is kept to zero volt ground level. The sensing electrode is also kept at ground level by the transimpedance amplifier which means the total lower plane of the sensor is at ground level and the field lines will only bend at the edges of the guard ring and not at the sense electrode.

The last item of errors is caused by the extremely small measurement current levels as mentioned before. This makes this sensor type extremely sensitive for interfering electromagnetic fields even when only the excitation

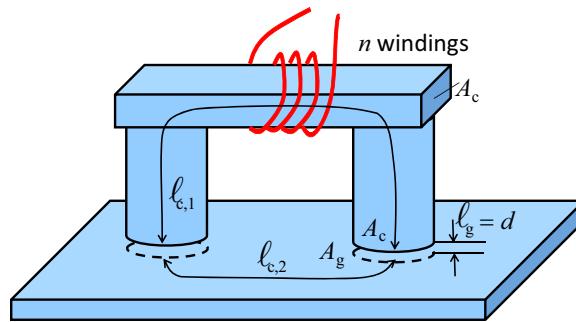


**Figure 8.38:** By using two electrodes and a guard ring, the distance to a moving conducting object can be measured without the need to directly contact the moving part. In this case the guard ring reduces the current that flows directly over the capacitance between the transmitting electrode and the sensing electrode.

frequency is filtered out with a selective filter. In case this sensor is applied in a high bandwidth real time feedback controlled system, all measures for shielding must be applied, while avoiding loops in the wiring. Furthermore, the sensing electronics should be located as close as possible to the sensor. With extreme critical applications this might even require measures for cooling the heat that relates to the often high power consumption of the required low noise amplifiers in order to avoid thermal errors.

### 8.6.2.3 Sensing to conductive moving plate

The guard ring method can in principle also be applied in a differential sensor to achieve linearity but in some cases it is even necessary to measure the distance to an object that can not be supplied directly with any electric current by means of a connected wire. As an example one can think of a continuously rotating shaft or a fast moving long-range linear positioning system. In that case it is possible to use a third electrode at the stationary part as shown in Figure 8.38. In this configuration a series of two capacitive sensors is obtained that both depend on the distance to the moving part. When the total active surface area is unchanged, the capacitance of this sensor is about a quarter of a normal sensor because the surface of the electrodes is half and the distance is passed twice. This implies a very small working distance in the order of 100  $\mu\text{m}$  or less for real precision measurements. In this case the guard ring is used to separate the transmitting electrode from the sensing electrode. This will reduce the currents that directly flow between both electrodes. When the transmitting electrode is much larger than the sensing electrode the field on the sensing electrode will determine the total current and in that case the more homogeneous



**Figure 8.39:** Distance and velocity sensor based on the variable reluctance of an air gap.

field at the sensing electrode will be also useful.

Also this sensor is used with virtual grounding by a transimpedance amplifier.

As a conclusion, a capacitive sensor has the ability to give information about the relative position of an object without any moving parts that are susceptible to wear. This means that these sensors can be used where a high reliability over an extended period of time is required.

### 8.6.3 Inductive position sensors

In Section 5.1.2 of Chapter 5 the inductor was introduced, consisting of a coil wound with  $n$  windings around a ferromagnetic core creating a magnetic flux according to the following equation for a homogeneous core:

$$\Phi_w = \frac{nI}{\mathfrak{R}} = nI \frac{A_c \mu_0 \mu_r}{\ell_c} \quad (8.43)$$

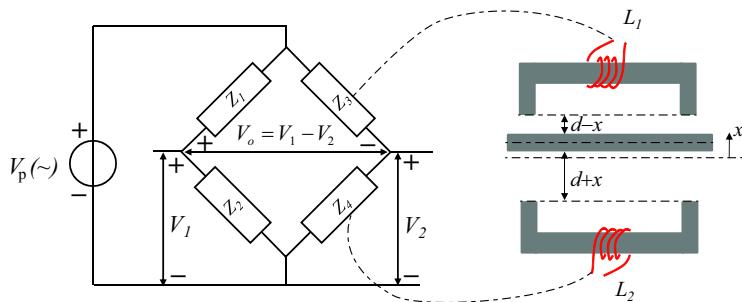
When the magnetic path consists of both a ferromagnetic core and an air gap, the total reluctance is the sum of the respective reluctances. For the situation as shown in Figure 8.39 the total reluctance equals:

$$\mathfrak{R} = \frac{\ell_c}{A_c \mu_0 \mu_r} + \frac{\ell_g}{A_g \mu_0} = \mathfrak{R}_0 + c_1 d \quad \text{with: } c_1 = \frac{2}{A_g \mu_0} \quad (8.44)$$

Note that the flux has to pass the air gap twice and that  $\ell_c$  consists of two parts, one part is the core with the windings and one part is the return path to the object of which the distance  $d$  is measured.

The self inductance was defined as follows:

$$L = \frac{\Phi_{w,t}}{I} = \frac{n\Phi_w}{I} = \frac{n^2}{\mathfrak{R}} \quad (8.45)$$



**Figure 8.40:** A differential reluctance sensor has a linear relation between the displacement  $x$  and the voltage  $V_o$ .

where  $\Phi_{w,t}$  equals the sum of the flux  $\Phi_w$  of each winding. It is further assumed that  $\Phi_w$  is equal for all windings. Using this definition, the self inductance becomes:

$$L = \frac{n^2}{\Re_0 + c_1 d} = \frac{L_0}{1 + c_2 d} \quad \text{with: } c_2 = \frac{c_1}{\Re_0} \quad \text{and: } L_0 = \frac{n^2}{\Re_0} \quad (8.46)$$

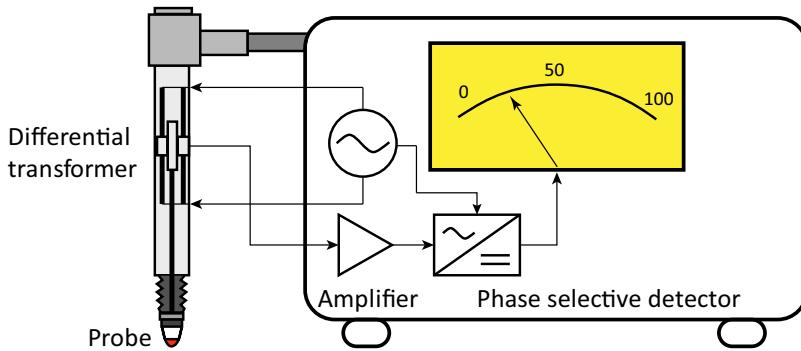
The voltage over a self inductor is proportional to the change of the total flux  $\Phi_{w,t}$  over time. This leads to the following relation:

$$V = \frac{d\Phi_{w,t}}{dt} = \frac{d}{dt} LI = L \frac{dI}{dt} + I \frac{dL}{dt} \quad (8.47)$$

This means that an inductive sensor can use two sensing principles: In case of an AC current through the coil the voltage is directly related to the position. With a DC current the voltage is proportional to the relative velocity. This second principle is hardly used for measurement as using permanent magnets is far more efficient to detect a relative velocity in a coil as will be explained later. For that reason this principle will not be pursued any further.

The first term however is important as the self inductance is inversely proportional to the distance, similar to a single capacitive sensor. The related non-linearity can be solved in the same way as with a capacitive sensor by using a Wheatstone bridge and two sensors in a balanced bridge as shown in Figure 8.40. The impedance of an inductor is proportional to the frequency ( $Z = j\omega L$ ). For that reason the other branch of the Wheatstone bridge is taken as with the capacitive sensor in order to get a positive output voltage with a movement in the positive  $x$ -direction. The voltage can be calculated with Equation (8.46) in the same way as with the differential capacitive sensor:

$$L_1 = \frac{L_0}{1 + c_2(d - x)}, \quad L_2 = \frac{L_0}{1 + c_2(d + x)} \quad (8.48)$$



**Figure 8.41:** A Linear Variable Differential Transformer can measure very accurately small displacements around the centre position.

The voltage  $V_o = V_1 - V_2$  becomes, when applying the rules of a voltage divider:

$$V_o = V_p \left( \frac{Z_2}{Z_1 + Z_2} - \frac{Z_4}{Z_3 + Z_4} \right) \quad (8.49)$$

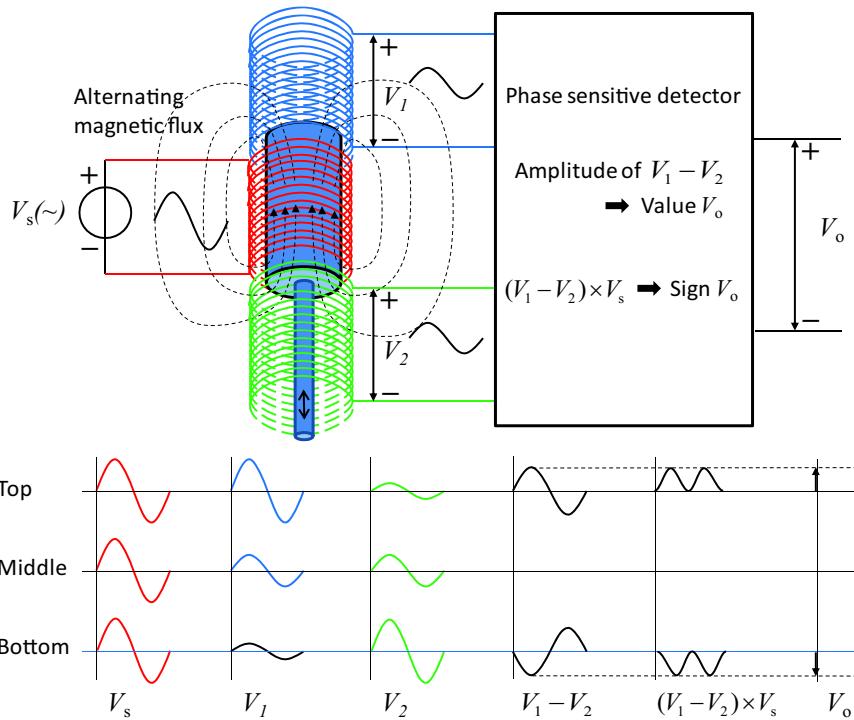
If  $Z_2 = Z_1$ , being just simple resistors in most cases and replacing  $Z_3$  and  $Z_4$  by the impedance of respectively  $L_1$  and  $L_2$ , the following equation gives the voltage as function of the displacement:

$$\begin{aligned} V_o &= V_p \left( \frac{1}{2} - \frac{j\omega L_2}{j\omega L_1 + j\omega L_2} \right) = V_p \left( \frac{1}{2} - \frac{1}{1 + c_2(d + x)} \right) = \\ &= V_p \left( \frac{1}{2} - \frac{1 + c_2d - c_2x}{2(1 + c_2d)} \right) = V_p \frac{c_2x}{2(1 + c_2d)} \end{aligned} \quad (8.50)$$

Also this relation is linear and frequency independent.

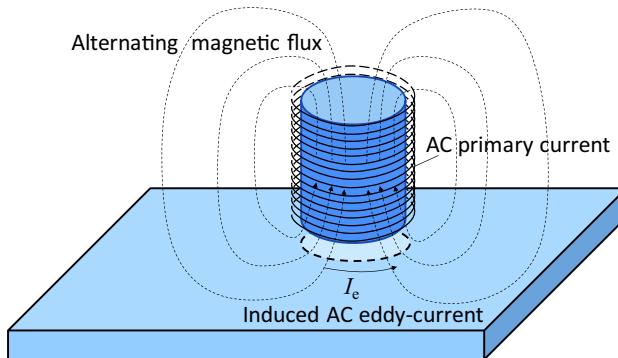
### 8.6.3.1 Linear variable differential transformer

In laboratory measurements and workshops the Linear Variable Differential Transformer (LVDT) as shown in Figure 8.41 is a frequently used instrument for measuring small displacements. It is based on the principle of a coupled alternating magnetic field like with the electrical transformer as was presented in Section 5.5. In Figure 8.42 the working principle is shown in more detail. The transformer consists of three coils where the middle coil is supplied with an AC voltage. The two secondary coils are positioned on both sides in line with the primary coil. Other than with the normal



**Figure 8.42:** Signals in an LVDT. The secondary coils in an LVDT partly share the flux of the primary coil. In the mid position the induced voltages are equal. A position change of the core will increase the flux at one coil while decreasing the voltage in the other. The difference voltage is then proportional to the displacement while the phase corresponds with the direction.

transformer the core is not closed but consists only of a small ferromagnetic rod with about the same size as the primary coil. In the mid position the ferromagnetic rod is located just inside the primary coil. The generated magnetic flux from the primary voltage will only partly be coupled to the secondary coils. The induced voltage in these coils therefore is small and equal for both coils. The subtraction of these voltages gives a value of 0 V, corresponding with this mid-position. When however the core is moved in the direction of either one of the secondary coils, the coupling with that coil will increase while the coupling with the other coil will decrease with a corresponding increase and decrease of the induced voltage. In that case the subtraction of both voltages results in a voltage with an amplitude that is proportional to the displacement. Depending on the direction, whether



**Figure 8.43:** Measuring the distance to a subject by means of an eddy-current sensor. The magnetic field induces a voltage over the conductive object. This creates a current that will suppress the magnetic field resulting in an increased current in the primary coil.

moving towards coil A or B, the phase of the resulting signal will be inverted or non-inverted giving the direction of the position. By using a phase selective detector a DC voltage is obtained with a sign that corresponds to the direction of the displacement. Phase selective detection can be achieved by taking the sign of the signal obtained by multiplying the input with the output and use the rectified and low-pass filtered output signal  $V_1 - V_2$  to get the amplitude. This principle was also explained in Section 8.4.2.2 regarding synchronous demodulation of an amplitude modulated signal.

### 8.6.3.2 Eddy-current sensors

As a last example of an inductive proximity detector the eddy-current sensor is presented here as it is frequently used in industrial systems to detect the proximity to an electrically conducting object. Like with the LVDT, its working principle is based on the interaction of two windings by their coupled magnetic field however in this case one of the windings is the object itself. When looking at Figure 8.43 the sensor consists of a single primary coil wound around a ferromagnetic open core. This primary coil is supplied with an AC voltage. The resulting magnetic field exits the core on one end and enters it at the other. As soon as the sensor approaches the electrically conducting surface of an object, the alternating magnetic field will induce an alternating voltage in this surface comparable with the secondary winding in a transformer. This voltage is however short circuited as this secondary winding is a solid piece of metal resulting in an eddy-current that will

suppress the magnetic field. This causes the current in the primary coil to increase just like with a regular transformer and by measuring this current, a value for the distance to an object is obtained. Information on the distance is available both in the amplitude and in the phase of the primary current relative to the primary voltage which is explained as follows. At a large distance from the conductive object the coil acts as an ideal self inductance giving a current with a low amplitude that is  $90^\circ$  delayed. As soon as the distance is smaller the eddy-current will induce a real primary current that adds to the imaginary current level because the eddy-current is in phase with the secondary voltage and consequently with the magnetic flux and the primary voltage. This also means that the consumed power of the sensor increases and the total current will be less than  $90^\circ$  out of phase.

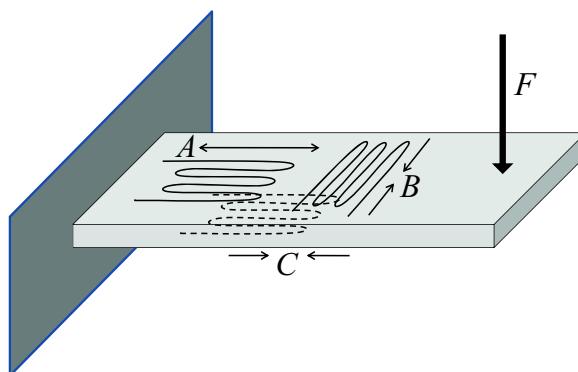
Eddy-current sensors are not very precise but measurements with  $\mu\text{m}$  accuracy around a fixed setpoint are possible like in magnetic bearings.

## 8.7 Dynamic measurements of mechanical quantities

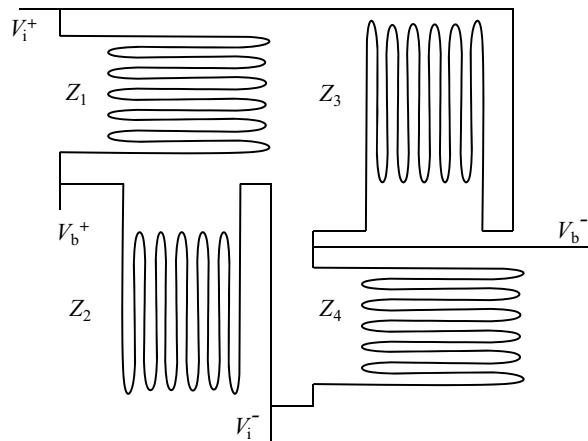
Mechatronic systems are ruled by dynamic, time varying signals that are determined by forces acting on the system. The measurement of these forces and the related velocity and acceleration of a positioning system is most important in controlling a positioning system for all states. Also in the dynamic modal analysis of a mechanical part inside a mechatronic system these dynamic sensors are frequently applied.

### 8.7.1 Measurement of force and strain

Force is a physical concept that is only indirectly observed at two different phenomena, strain and acceleration. The force we “feel” as a human being is the result of strain as our body deforms under the influence of a external force and we need to control our muscles to compensate that deformation. Generally, even with acceleration, force is measured by measuring the strain in a material that is exposed to a force. The strain gage has become the standard method for this measurement but lately also a very sophisticated modern optical method has been developed, using a *Fibre Bragg Grating* that will be presented after the strain gage.



**Figure 8.44:** A force sensing bending beam with different locations for the strain gages. The gage at location A is elongated while the gages at locations B and C are compressed when applying a force F in the downward direction.



**Figure 8.45:** A full bridge strain gage consisting of 4 strain sensitive resistors.

### 8.7.1.1 Strain gages

A strain gage is a piece of thin conductive material that changes its resistance as function of a change of its dimensions. An elongation in the direction of the current flow will increase the resistance both by the longer trajectory and by the corresponding contraction of its diameter by the stretching effect. These sensors are mainly used in non-destructive measurements with very limited strain levels. As a consequence the resistive change is also very limited in the order of less than 1 %.

A strain gage is a passive sensor and needs an external source of electricity to function. As explained before, the Wheatstone bridge is the best choice for this kind of sensor and in the following the Wheatstone bridge is taken as example. It was shown that the temperature effect in the Wheatstone bridge can be minimised by using a thermally balanced branch with two sensors with an opposite sign in the sensitivity. In Figure 8.44 it is shown how this can be realised with strain gages by mounting the second strain gage close to the first strain gage on a location that either shows an opposite strain like the reverse side of the bending beam or on the same side but orthogonal to the first strain gage where the poisson ratio of the material will cause the surface to show an opposite strain.

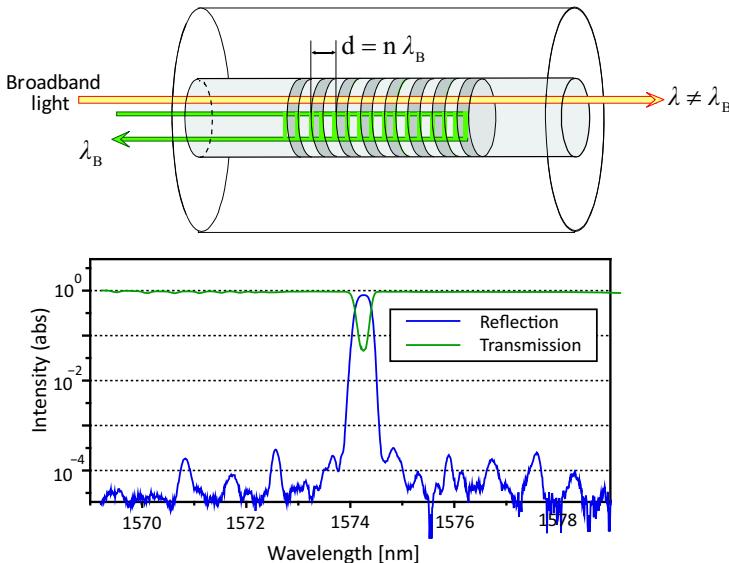
The most optimal configuration of a strain gage where both temperature and source noise is minimised relative to the sensitivity is the full Wheatstone bridge with four strain gages. This can be realised on one monolithic part when using the Poisson ratio effect according to the configuration of

**Table 8.1:** K factors of strain gage materials.

Material	K factor
Platinum-Iridium (Pt 95 %, Ir 5 %)	5.1
Platinum-Tungsten (Pt 92 %, W 8 %)	4.0
Isoelastic (Fe 55.5 %, Ni 36 % Cr 8 %, Mn 0.5 %)	3.6
Constantan / Advance / Copel (Ni 45 %, Cu 55 %)	2.1
Nichrome V (Ni 80 %, Cr 20 %)	2.1
Karma (Ni 74 %, Cr 20 %, Al 3 %, Fe 3 %)	2.0
Advance (Cu 54 %, Mn 1 %, Ni 44 %) <sup>7</sup>	2.0
Armour D (Fe 70 %, Cr 20 %, Al 10 %)	2.0
Monel (Ni 67 %, Cu 33 %)	1.9
Manganin (Cu 84 %, Mn 12 %, Ni 4 %)	0.47
N or P type Silicon	120-175!

Figure 8.45. Only 4 connections are needed to the measurement electronics. This configuration is frequently used in many force sensors (load-cells) and is proven to be very reliable.

Table 8.1 shows several materials that are used in strain gages, each with a different strain sensitivity. While  $K$  for most metals ranges between 0.5 and 5 the piezoresistivity effect in semiconductors gives a sensitivity that far outweighs the simple straightforward effect by the volumetric changes of a material under stress like in metals. This effect is due to the special conduction mechanism in semiconductors with holes and electrons which is also influenced by strain resulting in gage factors of 100 or more. This has enabled its use in fully integrated MEMS accelerometers that are frequently used in cameras and cell-phones for motion detection and sensing of the gravitational forces of the earth in order to detect the orientation. With these materials the full bridge configuration is even more necessary because of the also inherent very large temperature sensitivity of Silicon.



**Figure 8.46:** A fibre Bragg grating is created inside an optical fibre by means of an area with a longitudinal periodic variation of the index of refraction. A standing wave will occur at a wavelength  $\lambda_B$  of which an integer number just fits inside the grating period  $d$ . The light corresponding with that standing wave is reflected.

(courtesy of Technobis Fibre Technologies)

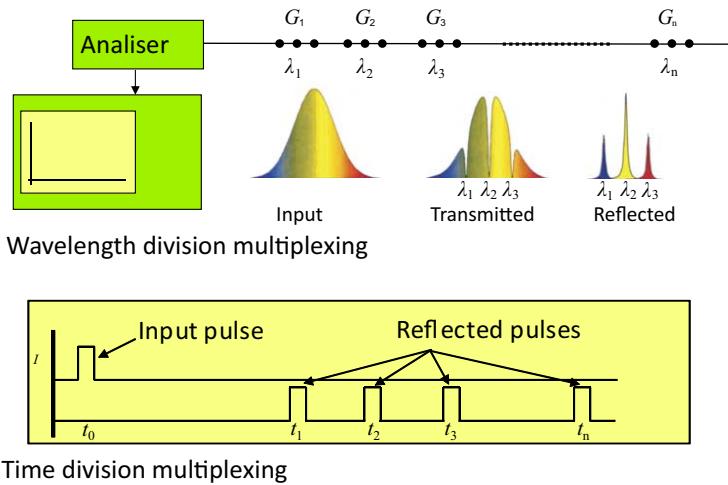
### 8.7.1.2 Fibre Bragg grating strain measurement

The fibre Bragg grating principle is discovered by the British physicist William Lawrence Bragg (1890 – 1971) by research on X-ray diffraction at crystalline materials, for which he won the Nobel prize. The observed effect, the reflection and refraction of certain wavelengths in non homogeneous material, is closely related to the multiple Fabry-Perot interferometer effect of optical coatings as explained in Chapter 7.

An optical fibre Bragg grating is created inside a transparent optical fibre by introducing small areas of approximately 3 mm with a periodic longitudinal variation of the index of refraction. When the periodicity is constant, the grating will cause a reflection of those light wavelengths  $\lambda_B$  that fit in the grating period  $d$  according to the expression  $d = n\lambda_B$ , where  $n$  is an integer and in most applications equal to one.

The working principle is illustrated in Figure 8.46.

When light from wide-band light source is inserted in the fibre, only the light



**Figure 8.47:** A fibre Bragg grating can be used to measure strain at long distances simultaneously at different locations on the fibre, distinguished by the wavelength. Also the temporal response of the different reflections gives information of the strain.  
 (courtesy of Technobis Fibre Technologies)

with a wavelength equal to  $\lambda_B$  will be reflected. An elongation of the fibre will result in a change in the reflected frequency because of this relation, which makes this frequency a measure for the strain of the fibre. By locating the fibre on or inside a solid object, the strain can be determined at the location of the grating.

In principle the grating can be located at a very long distance, but even more interesting is the fact that more than one grating can be applied on one fibre, as shown in Figure 8.47. With *Wavelength Division Multiplexing* (WDM) the different gratings are distinguished by giving them each a slightly different grating period as long as the spectral range of the frequencies that fit in each grating period do not have an overlap.

With some real numbers this becomes more clear.

In practice a light emitting diode (LED) can be used as a light source. In this example the LED has a wavelength of 850 nm and a spectral bandwidth of 50 nm. This bandwidth is divided in ten equal frequency ranges separated with 5 nm spectral difference with the other areas. Rejecting the outer areas with a lower irradiance results in practice in eight different sensing areas per fibre.

The grating period is  $\approx 1 \mu\text{m}$  due to the average refractive index of the fibre giving 3000 periods in the 3 mm measurement area.

The frequency spectrum is measured by means of a diffraction grating where the wavelength determines the diffraction angle. The diffracted beam exposes a CCD photo detector at a fixed distance from the grating and the position of incidence of the diffracted light is measured. This system has a maximum resolution of a wavelength change of one picometre, corresponding with 1  $\mu\text{m}$  per metre strain while the repeatability is better than three micro-strain.

The electronics are so fast that the dynamic bandwidth equals 20 kHz with a fixed latency of 50  $\mu\text{s}$ .

Next to this spectral difference, the distance between the gratings can be measured by determining the return time of the reflections by a pulsed light source. This is called *Time Division Multiplexing* (TDM).

Like with almost all sensing methods also this principle suffers from temperature as an interfering error source. The index of refraction of any material is temperature dependent to a certain degree but it can be compensated and even used by combining different fibers with different optical and thermal properties and that enables to measure both temperature and strain simultaneously.

With both WDM and TDM methods combined, large objects like the wings of an airplane, the vanes of a windmill and the cables of a hanging bridge can be measured, but also for measuring less large object this method offers ample possibilities.

## 8.7.2 Velocity measurement

The measurement of velocity has long relied on the interaction of the magnetic field of a permanent magnet with a moving conductive element. Before the onset of the electronic era the standard speedometer in cars and motorbikes all worked according to the principle that a moving conductive disk inside a magnetic field experiences a force that is proportional to the velocity.

Presently still the majority of cars are equipped with speedometers with needle pointers but they are no longer driven by moving disks in a magnetic field but by a moving-coil Lorentz type rotating actuator that is controlled by electronics that receive the velocity information from different sensors. These sensors often consist of systems with an incremental character, resulting in a frequency proportional to the velocity.

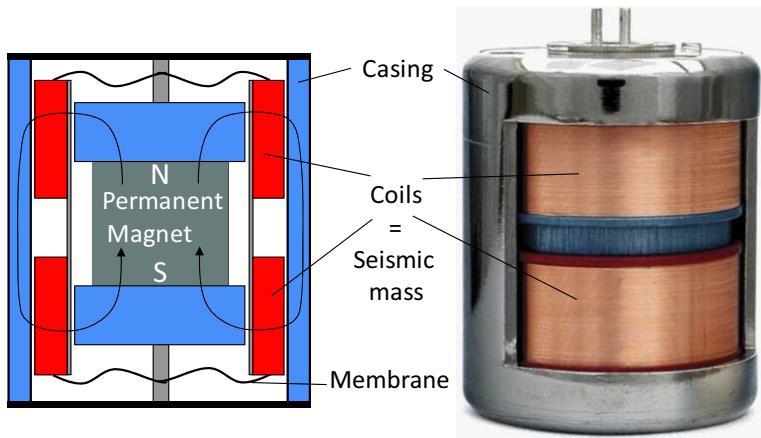
Incremental velocity sensors can be based on many principles. One example is using a light source, a photo sensor and a wheel with holes that intermittently blocks the light as function of the angle.

Another example is a rotating permanent magnet that induces an AC voltage in a stationary coil with a frequency and amplitude that is proportional to the speed according to the law of Faraday.

These principles will be presented more in depth in Section 8.8 with long range optical encoders and this subsection will focus on the velocity sensing that is needed in feedback systems to control damping.

It is often not necessary to use separate velocity sensors in precision positioning systems for damping in a full state feedback configuration. In most cases damping is achieved with PD-control by differentiating the position measurement signal as was explained in Chapter 4. In certain situations, however, it is better to use a separate sensor, especially when the reference for the damping action is undefined. This is the case with active vibration isolation systems like will be presented in Chapter 9. In those systems a damper connected between the isolated part and the vibrating environment introduces an increase in the transmissibility of the vibrations at higher frequencies. By using an active control system with a low stiffness actuator and a velocity sensor, damping can be created without an increase of the transmissibility as long as the sensor measures the velocity relative to a reference that is completely free from vibrations.

Such a quiet reference is realised with an *inertial velocity sensor* like the *geophone* that was originally designed to measure low-frequency vibrations like earthquakes (Geo-phone = sound of earth).



**Figure 8.48:** A geophone has a moving coil system supported by a compliant membrane around a permanent magnet that induces a magnetic flux through the coil. The movement of the coil will create an electro-motive force, proportional to the velocity of the movement.  
 (©Aaronia AG, [www.aaronia.com](http://www.aaronia.com))

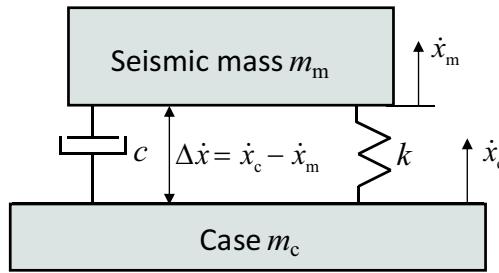
### 8.7.2.1 Geophone

The operating principle of a geophone is shown in Figure 8.48. A moving mass, also called *seismic mass*, consists of a coil in two sections that is supported inside a round ferromagnetic casing by means of a very compliant spring membrane, resulting in a very low resonance frequency in the order of 1 – 10 Hz.

The coil moves inside a magnetic field that is created by a permanent magnet inside the coil, connected to the casing. The magnetic field enters one of the coil sections from the inside and returns via the casing through the other section from the outside. The coil sections are wound in opposite directions and as a result their induced voltage will have the same sign. When added, the total voltage from the coil is proportional to its relative velocity to the stationary magnetic field, according to Faraday's law.

A geophone is used to measure velocity vibrations of an object by mounting it to the moving object. This means that the output voltage is determined by the transfer function from the velocity of the casing to the velocity difference between the casing and the coils as the latter is proportional to the voltage of the sensor.

It can be reasoned that under static conditions, at a frequency below the first resonance frequency, the seismic mass of the coils will follow the casing



**Figure 8.49:** The dynamic model of a geophone as used for deriving the transfer function.

and no voltage will be induced. Above the resonance frequency the seismic mass will decouple and its amplitude will rapidly drop resulting in a velocity difference and a measurable voltage. With the standard motion equations, the influence of the internal damping on this system can be determined. This damping is caused by eddy-currents in the coil former that often is made from solid copper for increased mass and behaves like a short circuited coil.

The relation between the velocity difference  $\Delta\dot{x} = \dot{x}_c - \dot{x}_m$  and the velocity of the casing  $\dot{x}_c$  is defined by the following equation using the rigid body model from Figure 8.49:

$$\frac{\Delta\dot{x}}{\dot{x}_c} = \frac{\dot{x}_c - \dot{x}_m}{\dot{x}_c} = \frac{\frac{d}{dt}(x_c - x_m)}{\frac{d}{dt}x_c} = \frac{s(x_c - x_m)}{sx_c} = \frac{x_c - x_m}{x_c} \quad (8.51)$$

With the second law of Newton the force balance on the moving mass  $m_m$  is defined:

$$m_m s^2 x_m = c(sx_c - sx_m) + kx_c - kx_m = (cs + k)(x_c - x_m) \quad (8.52)$$

From the theory on transmissibility from Chapter 3 it is known that:

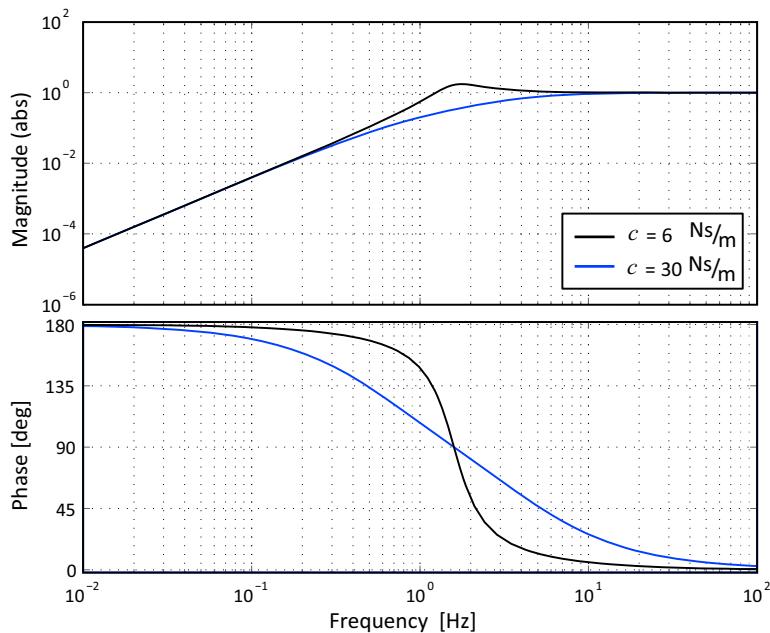
$$x_m = x_c \frac{cs + k}{m_m s^2 + cs + k} \quad (8.53)$$

Which combines with the previous expression into:

$$m_m s^2 x_c \frac{cs + k}{m_m s^2 + cs + k} = (cs + k)(x_c - x_m) \quad (8.54)$$

And leads to the following concluding equation:

$$\frac{\Delta\dot{x}}{\dot{x}_c} = \frac{\dot{x}_c - \dot{x}_m}{\dot{x}_c} = \frac{x_c - x_m}{x_c} = \frac{m_m s^2}{m_m s^2 + cs + k} \quad (8.55)$$



**Figure 8.50:** The Bode-plot of a geophone with a moving mass  $m_m = 1 \text{ kg}$ , a spring with a stiffness of  $k = 1 \cdot 10^2 \text{ N/m}$  and two values of the damping coefficient  $c = 6 \text{ Ns/m}$  (green) and  $c = 30 \text{ Ns/m}$  (blue) shows the effect of damping on phase and amplitude and the low-frequency bandwidth limitation.

When the voltage sensitivity of the moving coil equals  $V_g = c_g(\dot{x}_c - \dot{x}_m)$ , where  $c_g$  in  $\text{Vs/m}$  equals the sensitivity constant of the geophone, the total transfer function becomes:

$$\frac{V_g}{\dot{x}_c} = c_g \frac{\dot{x}_c - \dot{x}_m}{\dot{x}_c} = c_g \frac{m_m s^2}{m_m s^2 + cs + k} \quad (8.56)$$

This equation is a combination of the compliance response of a second-order mass-spring system with a second-order differentiator that act together as a second-order high-pass filter.

In this sensor the damping only influences the overshoot at the resonance frequency and can be chosen such that the overshoot is zero without impacting the high-frequency response.

In Figure 8.50 a Bode-plot of the transfer function of a typical geophone with a moving mass  $m_m = 1 \text{ [kg]}$  and a spring with a stiffness of  $k = 1 \cdot 10^2 \text{ [N/m]}$  is shown. The figure indicates that the overshoot at the resonance frequency is reduced with increasing damping but also the phase is changed.

Depending on the application, these two aspects can be optimised such that a flat response above the resonance frequency is obtained with an acceptable phase response.

It is clear from this example that the values of the mass and stiffness become impractical when the bandwidth is extended to even lower frequencies. Under gravity conditions this spring would already need to elongate with 0.1 m to compensate the static gravity-force on the mass, which means that a pre-stressed spring must be used.

This practical consideration limits the use of this geophone to a fixed orientation in respect to the earth gravity as otherwise a very long coil would be necessary, that would not be preferable because of the increased noise by the increased resistance.

In practice the LF-bandwidth limitation can be compensated by a suitable filter with integrator characteristics to increase the LF-gain, but this is limited because of the resulting amplification of LF noise (drift).

The main benefit of this geophone for the application in vibration isolation systems is the reference of the velocity to the inertial reference of a seismic mass as the inertial universe is by definition free from vibrations.

An alternative method to obtain this inertial velocity signal is based on using a sensor that even more directly refers to this inert reference by measuring acceleration. The velocity can then be derived by integration of the acceleration signal.

### 8.7.3 Accelerometers

Measurement of acceleration is in principle achieved by measuring force. The relation between acceleration and force is given by the well-known second law of Newton:

$$F = m \frac{dv}{dt} = m\dot{v} = ma \quad (8.57)$$

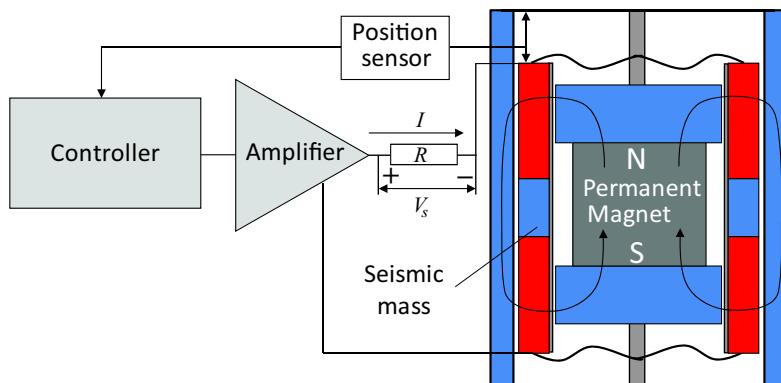
By using the force as an intermediate term, the acceleration is related to the distortion of the spring which creates the possibility to measure acceleration by using a combination of a seismic mass with a deforming support.

Three types of acceleration sensors or accelerometers will be presented based on this principle.

- The closed loop feedback accelerometer.
- The piezoelectric accelerometer.
- The MEMS accelerometer.

#### 8.7.3.1 Closed-loop feedback accelerometer

One way to measure a force is by compensating that force with the same force by means of an actuator. When the force to current ratio of that actuator is known, the current that is necessary for the compensating force is a



**Figure 8.51:** Closed loop feedback accelerometer. The controller suppresses the movement of the seismic mass relative to the casing with a current proportional to the acceleration.

reliable measure for this force. In Figure 8.51 this principle is used in the *closed-loop feedback accelerometer*. It is build with the same components as the geophone with the addition of a position sensor and a control system. In this configuration the coils function as a linear Lorentz actuator instead of a sensing element. The position sensor measures the position of the moving coils relative to the casing. This position sensor can be any reliable proximity detector as previously presented. The controller and the amplifier are tuned to keep the displacement as small as possible by supplying a current to the coils. This position feedback system creates an additional stiffness by the proportional gain. The necessary current for that action is proportional to the force on the mass due to the acceleration. By measuring the current, a value is obtained for the acceleration.

In Chapter 4 it was shown how such a position controller should be made in order to remain stable by placing the poles of the system on the right location in the complex plane with an additional D-control action. The system is a servo-feedback system as the controller has to control the seismic mass to follow the case just like with the optical pick-up unit of the CD player and the same PID-control principle can be used here.

Because of the fact that this servo controlled accelerometer is a zeroing measurement system, the position sensor inside the accelerometer only needs to be accurate around zero. This implies that a relatively simple sensor can be used with only a low level of the signal drift around zero when static accelerations over a longer period must be measured. A second benefit of this configuration is the additional virtual control-stiffness of the spring due to the feedback that reduces the influence of any non-linearity in the stiffness of the support structure in the frequency area where the closed-loop gain is high.

The feedback principle with a high gain at low frequencies makes this closed-loop feedback accelerometer especially suitable for the measurement of static or low-frequency accelerations like gravity and the slow motion of ships or air planes.

It is also a suitable option for low-frequency damping in vibration isolation systems when the sensor is combined with an integrator to obtain the inertial velocity. The complete sensor is however quite costly when compared to the two other accelerometers so the application of a closed-loop feedback accelerometer is limited to professional use at very low frequencies.

### 8.7.3.2 Piezoelectric accelerometer

In Section 5.6 of Chapter 5 the direct piezoelectric effect was explained. The internal charge distribution of a piezoelectric material will displace when deformed because of a special crystal structure with atoms that are constrained in the lattice. The charge displacement manifests itself as a voltage on the electrodes of the material.

Depending on the crystal orientation and the polarisation direction, the material is sensitive for shear strain or compressive strain and both principles are used.

Figure 8.52 shows three different configurations of a piezoelectric accelerometer.

The **shear design** is built with the seismic mass around the piezoelectric crystal. Similar to the shown example with one piezoelectric element also other versions exist of this principle with for instance an assembly of three flat piezoelectric plates in a triangular configuration. The main benefit of this configuration is its low sensitivity for residual strain in the base that can be caused by the mechanical mounting and by thermal induced deformations. The rotational symmetric configuration reduces also the sensitivity for accelerations in other directions.

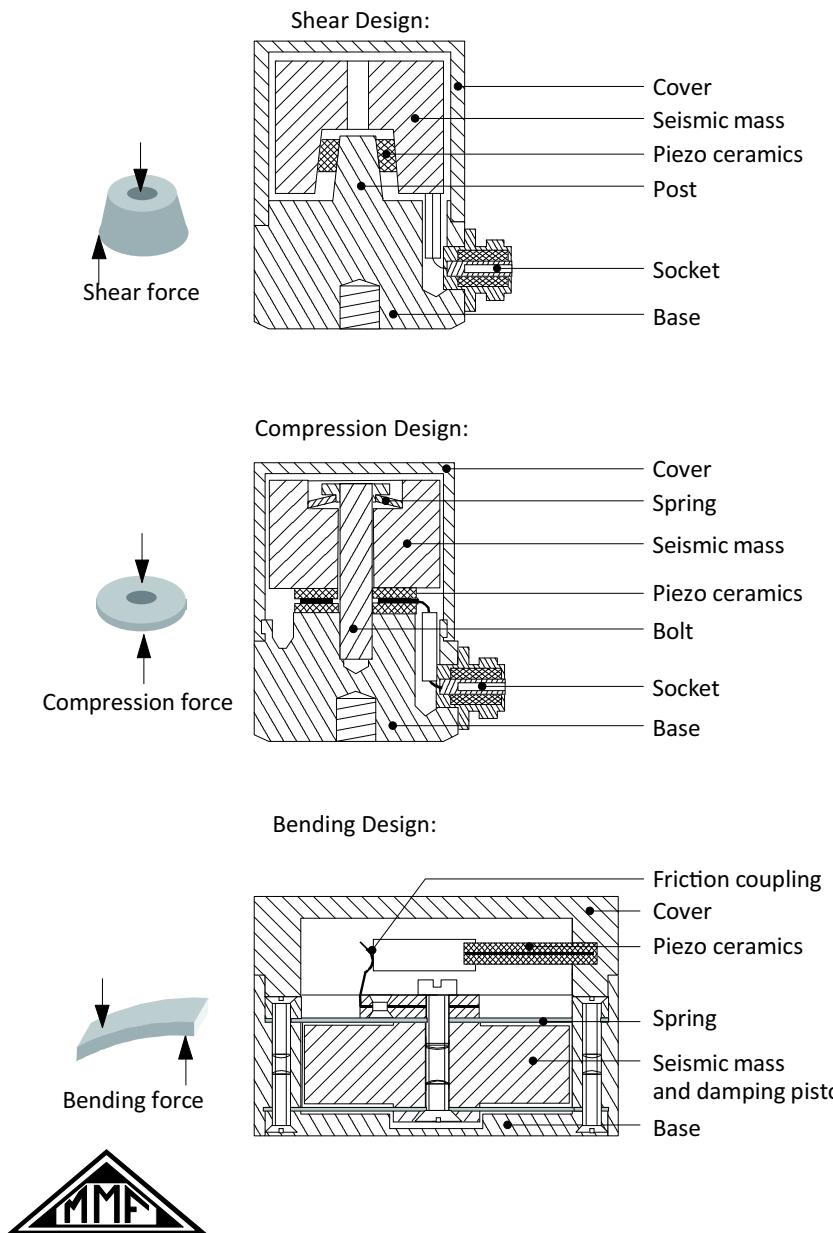
The drawback of the shear design is a more complicated assembly process, especially when a very small conical angle is needed. In assembly, excessive stress must be avoided, which requires careful handling.

The **compression design** is quite straightforward. The seismic mass is pressed onto a flat piezoelectric ring with a bolt of which the force is controlled by a flexible washer.

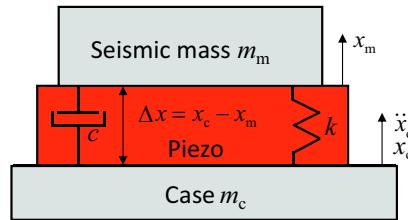
The compression design is more sensitive for thermal transients, base strain and accelerations in shear directions.

The **bending design** uses a special piezoelectric composition that consists of two layers, called a *bimorph*. A bending force from the seismic mass will create compressive strain in one layer and stretching strain in the other layer. Both layers are polarised in the opposite direction. As a result the voltages of both layers add together to one voltage between the two flat sides of the bimorph.

The main advantage of this principle is its high sensitivity-to-mass ratio. Its main drawback is its relatively low resonance frequency that limits the application to low frequencies as will become clear in the following section.



**Figure 8.52:** Three different mechanical designs of piezoelectric accelerometers. The shear design is the most modern version with better dynamic performance.  
(Courtesy of “Metra Mess- und Frequenztechnik Radebeul”)



**Figure 8.53:** The mechanical dynamic model of a piezoelectric accelerometer as used for deriving the transfer function. The voltage is proportional to the deformation  $\Delta x$  of the piezoelectric sensing element and relates to the acceleration  $\ddot{x}_c$  of the case.

### Mechanical frequency limitation

Mechanically a piezoelectric accelerometer behaves like a mass-spring system with a resonance frequency that is determined by the seismic mass and the stiffness of the piezoelectric crystal. The transfer function can be derived with the help of Figure 8.53. The voltage  $V_a$  of the accelerometer is proportional to the displacement  $\Delta x$  of the seismic mass that is caused by the acceleration  $a = \ddot{x}_c$  of the case.

$$\frac{V_a}{\ddot{x}_c} = c_v = \frac{c_d(x_c - x_m)}{\ddot{x}_c} = \frac{c_d(x_c - x_m)}{s^2 x_c} \quad (8.58)$$

where  $c_v$  in  $\text{Vs}^2/\text{m}$  equals the nominal voltage sensitivity of the accelerometer and  $c_d$  in  $\text{V/m}$  equals the voltage sensitivity of the piezoelectric crystal as function for its deformation.

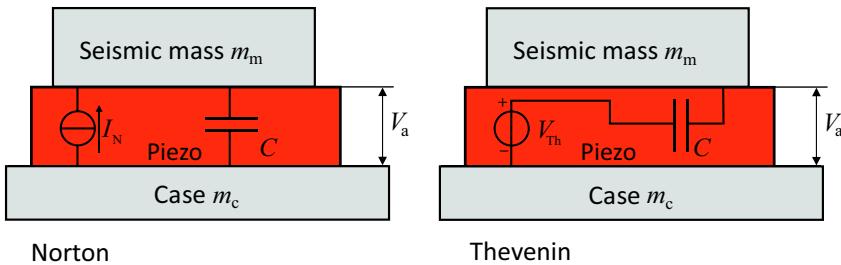
The equation is equal to Equation (8.56) of the geophone, divided by  $s^2$ . This means that the mechanical transfer function of a piezoelectric accelerometer is simply derived from that equation:

$$\frac{V_a}{\ddot{x}_c} = c_v = \frac{c_d(x_c - x_m)}{s^2 x_c} = \frac{c_d m_m}{m_m s^2 + c s + k} = \frac{c_d \frac{m_m}{k}}{\frac{m_m}{k} s^2 + \frac{c}{k} s + 1} \quad (8.59)$$

For low frequencies where  $s$  is very small the voltage sensitivity equals:

$$c_v = c_d \frac{m_m}{k} \quad (8.60)$$

The transfer function of Equation (8.59) is similar to the compliance response of a mass-spring system. Below the resonance frequency the output voltage is frequency independent. Around and above the resonance frequency the transfer function deviates from the constant value depending on



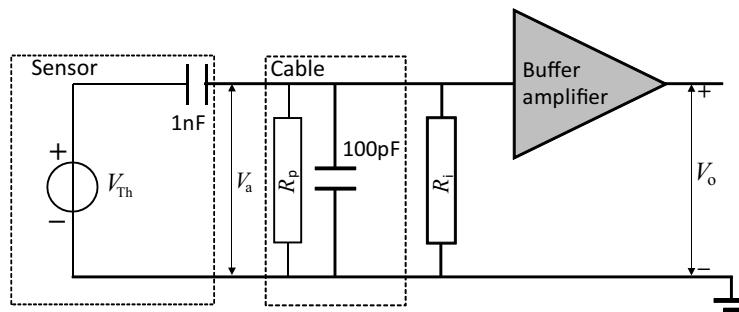
**Figure 8.54:** The electrical model of a piezoelectric accelerometer is basically a capacitor with a variable charge. This corresponds with the Norton equivalent model where the charge equals the integral of the frequency dependent current  $I_N$  over time. The Thevenin equivalent model is more practical to use because the voltage  $V_{Th}$  is not frequency dependent.

the damping. In practice the damping is quite low which means that these mechanical dynamics limit the high-frequency bandwidth of a piezoelectric accelerometer to around  $0.3 \times$  the resonance frequency. For the shear and compression design the support stiffness is rather high, giving practical values for the usable maximum frequency of several kHz, depending in the size. The bending design is more compliant. With the same size and mass as with the other design principles the application of the bending design is limited to high-sensitive low-frequency measurements until an approximate maximum of a few hundred Hertz.

For situations where only a very low mass is allowed at the measurement location, the bending design can still be a better alternative because with a very small size the resonance frequency can become acceptably high again.

### Electrical frequency limitation

Electrically a piezoelectric accelerometer behaves like a capacitor with a variable charge that is determined by the acceleration  $\ddot{x}_c$ . The change of charge is equivalent to a current which means that the sensor can be modelled according to Norton as a current source with a parallel capacitor. Unfortunately the current  $I_N$  is determined both by the acceleration amplitude and by the frequency. This is explained as follows: The location of the charge is determined by the force that is proportional to the acceleration. Only a change in the charge location will cause a current. This means that only a changing acceleration will cause the force, deformation and charge location



**Figure 8.55:** The parasitic capacitance and resistance of a cable and the input resistance of a buffer amplifier affect the sensitivity and frequency response of a piezoelectric sensor.

to change with a corresponding current.

With  $c_i$  being the charge sensitivity for acceleration, the Norton source current  $I_N$  becomes:

$$I_N = \frac{dQ}{dt} = c_i \frac{d\ddot{x}_c}{dt} = c_i s \ddot{x}_c = c_i j \omega \ddot{x}_c \quad [\text{A}] \quad (8.61)$$

This means that the Norton model is less practical due to this frequency dependence. The Thevenin equivalent voltage source  $V_{Th}$  can be derived from the Norton equivalent by taking the voltage of the sensor when no other load is applied than the internal capacitance:

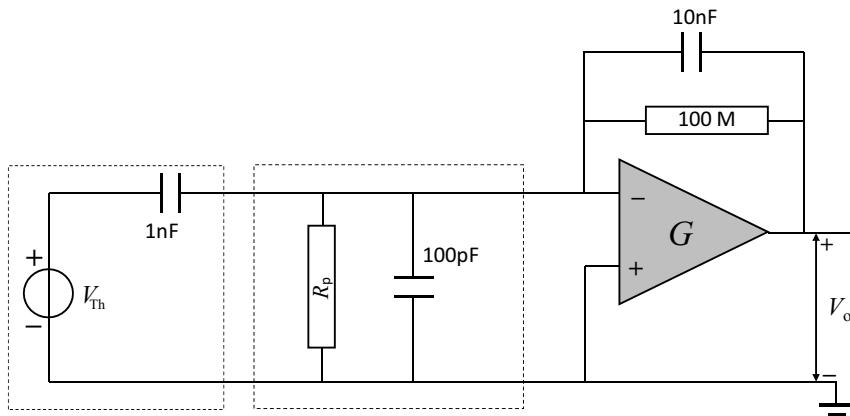
$$V_{Th} = I_N Z = \frac{c_i j \omega \ddot{x}_c}{j \omega C} = \frac{c_i \ddot{x}_c}{C} = c_v \ddot{x}_c \quad [\text{V}] \quad (8.62)$$

This is a frequency independent voltage and can be detected by a non-inverting amplifier. This amplifier would need to have a very high input impedance because the source impedance is capacitive. The high output voltage of a piezoelectric accelerometer allows the application of a unity-gain buffer amplifier as signal conditioning element when the capacitive nature of the sensor would be the only problem.

### Charge amplifier

A simple non-inverting buffer amplifier does not solve all problems as is indicated by the example of Figure 8.55 with some realistic values of the components.

When the sensor is connected to an amplifier, a well shielded coaxial cable is necessary to prevent external interference. This cable will have a parasitic



**Figure 8.56:** A piezoelectric accelerometer with a charge amplifier as signal conditioning element. The charge amplifier reduces the influence of the cable capacitance and other parasitic impedances by virtually grounding the output of the sensor.

capacitance that acts as a load to the sensor and gives a proportional reduction of the sensitivity, depending on the ratio between the cable capacitance and the sensor capacitance. Although this is a not frequency dependent attenuation, these values should be exactly known in order to take this attenuation into account with accurate measurements. This is only possible up to a certain level due to the temperature dependence of these values. Another problem is related to a possible parasitic resistance over the connector that might be caused by humidity in combination with salt residues due to manual handling.

Together with the input impedance of the amplifier and the capacitive source impedance of the sensor this circuit forms a first-order high-pass filter. When for this example the input resistance  $R_i$  equals  $50 \text{ M}\Omega$ , this low-frequency limit would be:

$$f_{LP} = \frac{\omega_{LP}}{2\pi} = \frac{1}{2\pi RC} = \frac{1}{2\pi \cdot 5 \cdot 10^7 \cdot 1.1 \cdot 10^{-9}} \approx 3 \quad [\text{Hz}] \quad (8.63)$$

because the total capacitance equals the impedance of the  $1 \text{ nF}$  and  $100 \text{ pF}$  capacitor in parallel.

A second problem is caused by the inherent non-linearity of a piezoelectric crystal.

As was shown with piezoelectric actuators these crystals exhibit a certain level of hysteresis between the voltage and the deformation. Fortunately this hysteresis is far less present in the relation between the charge and the

deformation as the hysteresis is caused by the capacitance. This means that an optimal way to detect the signal of a piezoelectric accelerometer would be to directly determine the charge. This can be achieved by “tapping” the current from the charge displacement by means of the charge amplifier of Figure 8.56, before it can cause a change in the voltage. This solution is especially useful because the charge amplifier also solves the first problem of the sensitivity for the cable capacitance and the parasitic resistance.

A charge amplifier is an operational amplifier in an inverting amplifier configuration with only a feedback impedance that is mainly determined by a capacitor. The sensor is directly connected to the inverting input of the operational amplifier.

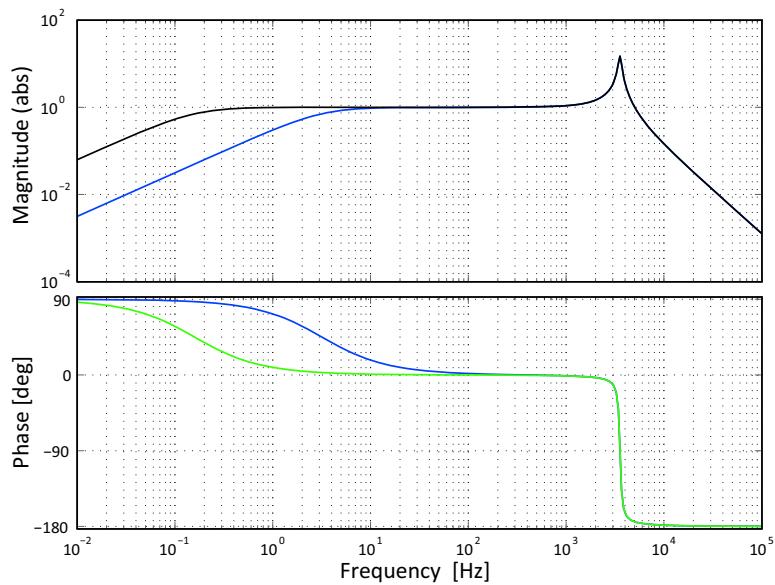
For the explanation of the principle first the effect of the very large  $100 \text{ M}\Omega$  feedback resistor can be neglected. The feedback capacitor for this example has a value of  $10 \text{ nF}$ . The rules of an operational amplifier from Chapter 6 are used to determine the gain of the inverting amplifier. The input impedance  $Z_1$  is equal to the source impedance of the sensor with a capacitor value of  $1 \text{ nF}$  and the feedback impedance  $Z_2$  equals the impedance of the  $10 \text{ nF}$  capacitor that was chosen for this example. The transfer function of this amplifier can now be written as follows

$$G = \frac{V_0}{V_{\text{Th}}} = -\frac{Z_2}{Z_1} = -\frac{\frac{1}{j\omega \cdot 10^{-8}}}{\frac{1}{j\omega \cdot 10^{-9}}} = -0.1 \quad (8.64)$$

Although this does not look impressive from a magnitude point of view, the effect of the parasitic impedances is cancelled as the amplifier will keep the voltage at the minus input equal to  $0 \text{ V}$ , being the voltage at the grounded plus input. This means that the sensor is virtually grounded and no voltage is present over the parasitic impedances, while all current of the sensor flows into the feedback impedance.

As this impedance is merely capacitive the voltage at the output is equal to the stored charge in the feedback capacitor, divided by its capacitance value. This charge is equal to the displaced charge of the piezoelectric accelerometer for which reason this configuration is called a charge amplifier. This might seem strange as the charge itself is not amplified and a “charge-to-voltage converter” would probably have been a better name.

The reduction of the voltage gain to  $0.1$  is the direct result of a design choice which is related to the maximum attainable value of the parallel resistor over the feedback capacitor. In theory this resistor is not needed but the bias current of the operational amplifier requires a non-infinite source



**Figure 8.57:** Typical bode-plot of a piezoelectric accelerometer with a seismic mass of  $20 \cdot 10^{-3}$  kg and a support stiffness of  $1 \cdot 10^7$  N/m showing both the HF bandwidth limitation by the mechanics and the LF limitation by the electronics. The amplitude plot is normalised to a value of one in the pass-band, the response of the blue line corresponds with a non-inverting buffer amplifier and the green line with the charge amplifier.

impedance at 0 Hz. The bias current level determines the maximum value of the parallel resistor together with the allowable DC offset error voltage at the output. Special amplifiers with MOSFET inputs are developed to achieve extremely low bias currents in the order of 1 pA which would give a DC error of 0.1 mV with the indicated  $100 \text{ M}\Omega$  resistor. Compared with the voltage amplifier situation parasitic impedances parallel to this resistor can more easily be avoided by careful assembly of the circuit, avoiding contamination in handling and when necessary, even a protective coating can be applied, to keep humidity away from the sensitive input of the amplifier. In this case the low-frequency limitation of the bandwidth is only determined by the feedback impedances resulting in the following:

$$f_{LP} = \frac{\omega_{LP}}{2\pi} = \frac{1}{2\pi RC} = \frac{1}{2\pi \cdot 1 \cdot 10^8 \cdot 10 \cdot 10^{-9}} \approx 0.16 \quad [\text{Hz}] \quad (8.65)$$

This is a factor 20 better than with the buffer amplifier. The sacrifice of the attenuation is easily solved by an additional voltage amplifier after the

charge amplifier. Furthermore, High-frequency voltage noise is no problem because of the capacitive feedback of the charge amplifier. The inherent low impedance for high frequencies results in a low gain for this kind of noise.

In Figure 8.57 a typical Bode-plot of a piezoelectric accelerometer with the two types of signal conditioning amplifiers is shown for comparison reasons. The data correspond with the examples of Figure 8.55 and Figure 8.56. The mechanical data of the corresponding sensor are a seismic mass of  $20 \cdot 10^{-3}$  kg and a support stiffness of  $1 \cdot 10^7$  N/m giving a resonance frequency of approximately 3.5 kHz. When using a charge amplifier, the maximum usable bandwidth of this example ranges from  $\approx 0.2 - 1000$  Hz.

Because of the inherent limitation of the bandwidth on both the HF and the LF side the piezoelectric accelerometer is used in situations where vibrations in specific frequency bands are investigated like in the dynamic modal analysis of mechanical structures.

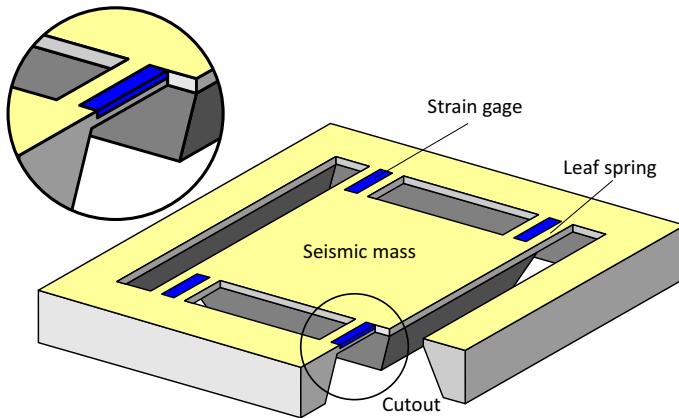
It is a frequently applied instrument in determining improvements in mechanics used in precision positioning systems like stages in wafer scanners. An additional benefit over other accelerometers is the very large dynamic range which is also due to the low HF noise level originating from its capacitive nature. The only noise that still can cause problems is low-frequency excess noise as that will be amplified by the feedback capacitor. Together with the low-frequency limitation of the sensing principle itself, this low-frequency noise is one of the problems when using this accelerator as velocity detector with an additional integration.

### 8.7.3.3 MEMS accelerometer

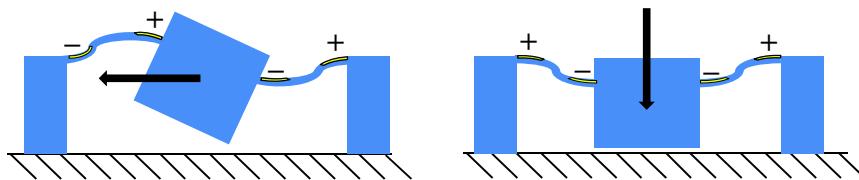
Micro Electro-Mechanical Systems or MEMS are based on the application of semiconductor production technology in the realisation of miniature mechanical designs.

By using materials like Silicon and Germanium that not only have very special electrical properties but also superior mechanical properties due to their mono crystalline structure, dynamic systems with for instance a very high Q (resonators and actuators) and an excellent linearity and sensitivity (sensors) can be realised. The possibility to directly integrate the signal conditioning electronics with the sensor itself on the same material guarantee a relatively low sensitivity for external magnetic and electrostatic fields.

For measuring purposes, the MEMS accelerometer, of which a very basic principle is shown in Figure 8.58, might be the most widely used MEMS



**Figure 8.58:** MEMS accelerometer that consists of a seismic mass supported by 4 leaf springs provided with semiconductor strain gages. In spite of the low mass this sensor has a very high sensitivity from 0Hz up till very high frequencies. (Courtesy of Fredric Creemer, DIMEs,TUD).



**Figure 8.59:** Vibrations in multiple directions can be measured with a MEMS accelerometer by using two strain gages per flexure.

device ever. It is applied in safety devices like the air bag for a car, in photographic cameras and Personal Digital Assistants (PDAs) to determine the orientation towards gravity.

Like with the other accelerometers, the MEMS accelerometer applies a seismic mass that is supported by a compliant mechanism, consisting of several flexure leaf springs. The position of the mass can be measured both with capacitive sensors and with integrated piezoresistive strain gages on the supporting springs. The capacitive sensing principle is frequently used in MEMS devices because of the simplicity. With comb structures the capacitance sensitivity can be increased and manufacturing these structures is relatively easy. With capacitive sensing linear accelerations can be detected in three orthogonal directions.

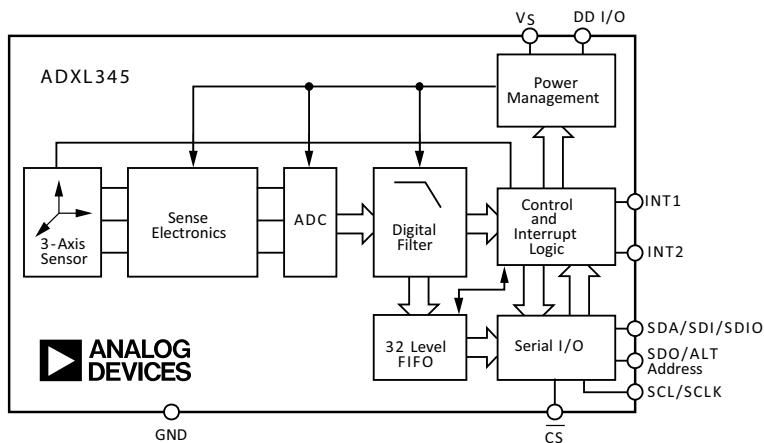
Measuring strain with integrated piezoresistive strain gages is an alternative that requires a bit more processing but it gives a very high sensitivity

and enables to sense even in six directions including rotational accelerations.

The principle is shown in Figure 8.59.

A representative example of a MEMS accelerometer is the ADXL345 from Analog Devices of which the functional diagram is shown in Figure 8.60. This fully integrated sensor works with the capacitive sensing principle and can measure accelerations in 3 directions to a maximum acceleration of  $160 \text{ m/s}^2$  with a resolution of  $0.04 \text{ m/s}^2$ , a non-linearity of 1 % of the full scale, an offset of less than  $1 \text{ m/s}^2$  and a maximum 13 bit wide data-rate of 3.2 kHz.

It is to be expected that these accuracy and noise levels will continuously improve over time which means that MEMS accelerometers will ultimately replace most other acceleration sensors because of their versatility, reliability, frequency range and low cost.



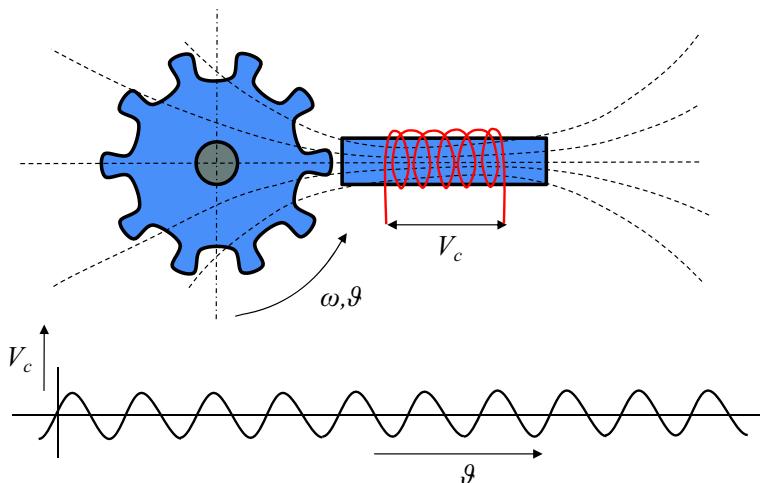
**Figure 8.60:** Functional diagram of the three directional MEMS accelerometer ADXL345 showing the full integration of the sensor, the signal conditioning and signal processing part that directly outputs the digital measurement data.

(Courtesy of Analog Devices).

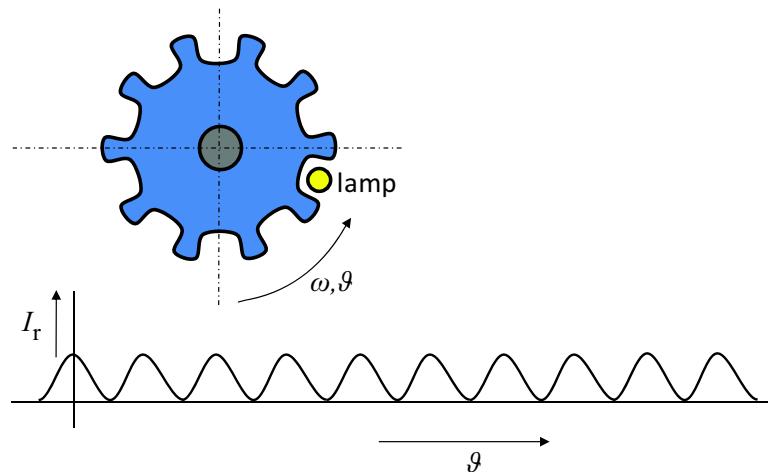
## 8.8 Optical long-range incremental position sensors

Analogue measurement of distances over a long range is limited by the signal to noise ratio of the applied sensor. A long range inherently conflicts with low noise that is needed for a sufficient resolution. For that reason long distances have always been measured in an incremental way for instance by counting the number of revolutions of a wheel or the number of stripes that were passed on a scale. In the age of electronics, incremental counting is not done anymore with rotating numbered disks interconnected by gear wheels but with digital logical circuits and microprocessors. When applying electronics, the only way to realise a long range high resolution position measurement system, is to provide a reliable signal with a sufficiently fine spaced spatial frequency.

This signal can be generated in different ways. Take for example Figure 8.61 with a permanent magnet inside a coil that will give a sinusoidal voltage as a function of the velocity and the angle of a toothed wheel that creates an alternating reluctance path for the magnetic field. This sinusoidal signal can be converted into a one bit periodic signal by means of a Schmitt trigger before counting the pulses. This principle however only works at a certain velocity, as at stand still no variable reluctance nor flux is observed. This



**Figure 8.61:** Incremental inductive rotation sensor giving a sinusoidal output voltage with a frequency depending on the rotation speed and the number of teeth.



**Figure 8.62:** Optical version of the incremental rotation sensor. The magnitude of the irradiance  $I_r$  from the lamp is modulated by the teeth of the rotating wheel.

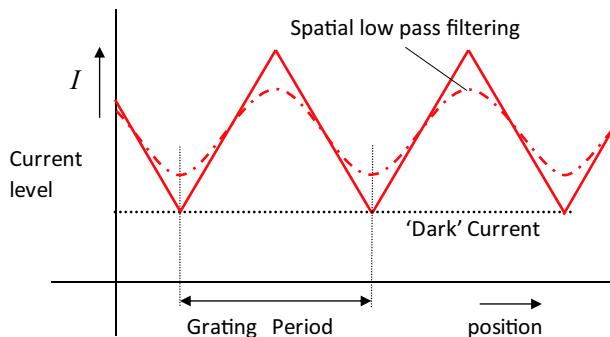
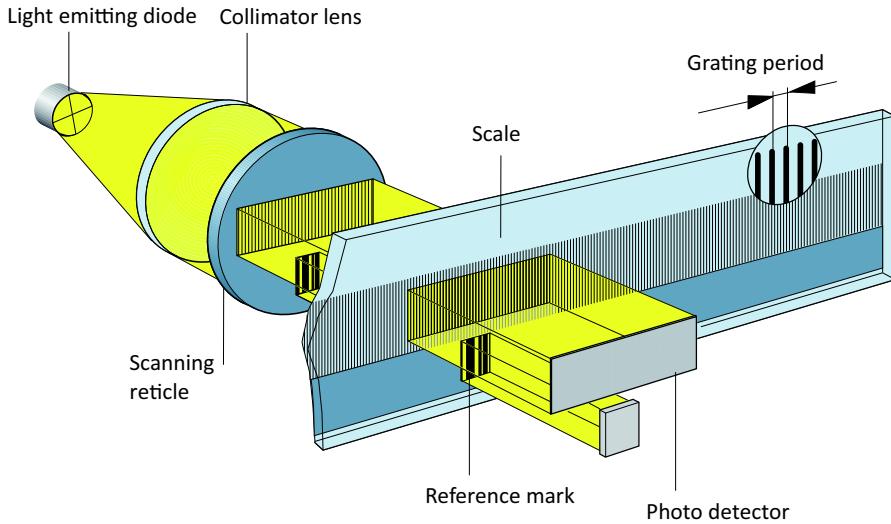
can be solved by using light instead of magnetism, as shown in Figure 8.62. By just using a sufficiently small lamp and a detector at the opposite sides of the rotating disk the detector will receive an alternating irradiance signal as function of the angle of rotation. This principle is refined in the optical encoders of the next section while the generation of an incremental signal by interference is used in the laser interferometers of the last section of this chapter. Both principles have presently achieved such high standards of measurement that they even can compete in resolution with many short range analogue sensors while still maintaining long measurement ranges,

### 8.8.1 Linear optical encoders

Optical encoders are widely applied both for angular and linear measurements. This section will restrict to the linear versions, because in precision mechatronic systems direct linear measurements are most commonly applied. Nevertheless, the same principles are applied in angular measurement systems.

The most basic example of a linear optical encoder is shown in Figure 8.63<sup>8</sup>. The working principle is originally based on the idea to compare the position

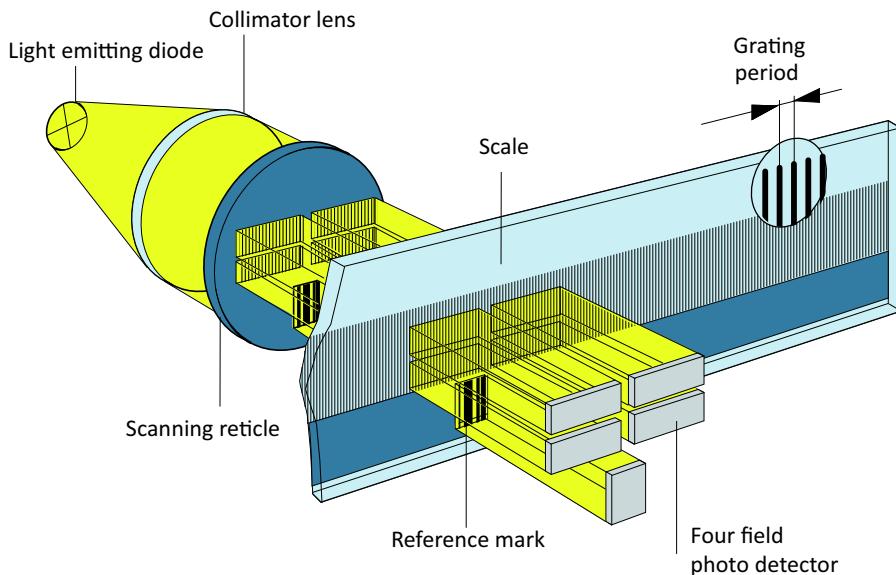
<sup>8</sup>This figure, together with several of the other figures and the background information in this section on optical encoders is graciously provided by **Heidenhain**, a leading company in high precision optical encoders.



**Figure 8.63:** Basic configuration of a linear optical encoder. A scanning reticle with a grating is illuminated with parallel rays of light. The shadow of the reticle is projected on a scale with an equal grating peiod connected to the moving object. A photo detector detects the resulting irradiance, that alters periodically as function of the displacement.

of one grating with well defined equidistant slits with the position of another grating by means of shadow projection.

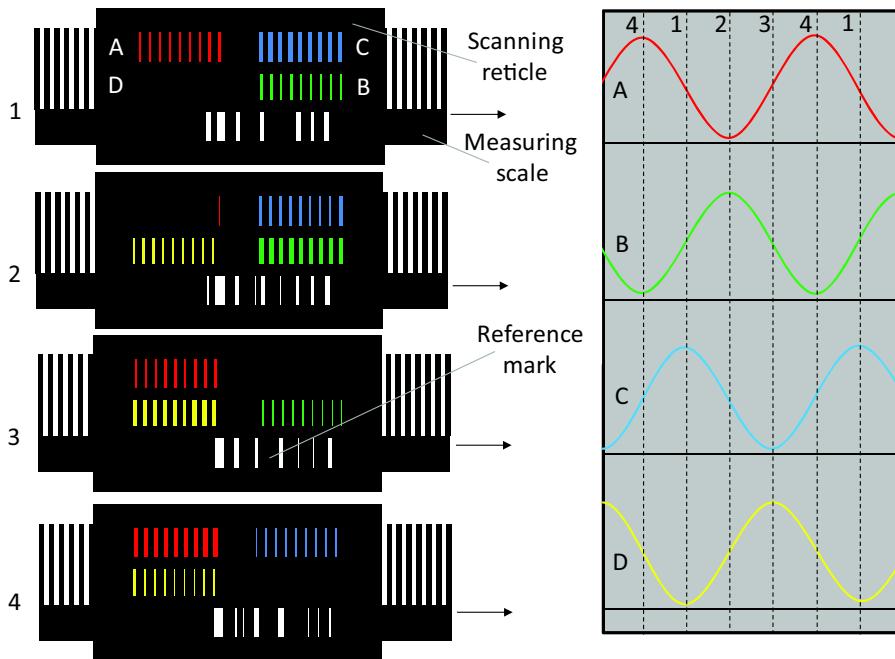
A parallel beam of light, originating from a light emitting diode and a collimator lens, is projected on the first grating. The shadow of this grating is projected on the second grating which has equal periodicity and is designed such that it obscures the light when shifted with half a period in respect to the first grating while transmitting the light when the gratings are in phase. This causes the irradiance to be modulated as function of the position



**Figure 8.64:** Four field incremental optical encoder giving robustness, directional and absolute position information with high resolution by interpolation.

with an essentially triangular spatial waveform. As will be explained later, a more sinusoidal spatial waveform is preferred. This can be created by spatial low-pass filtering of the triangular waveform. In Chapter 7 the effect of low resolution imaging optics was presented to be equivalent to low-pass spatial filtering. A shadow is only sharp under two conditions. Firstly the light must be created by a point source. Secondly the details need to be large relative to the wavelength of the light in order to prevent diffraction effects. With a larger surface of the light source a small grating period pattern becomes blurred, resulting in a more sinusoidal spatial waveform. With high resolution encoders this diffraction character determines the functionality, but for simple encoders this is only a side effect.

One typical property of incremental measurement systems is the relative character of the data. At a certain known position the counter is set to zero and, as long as all periods are counted during movement, the new position is determined relative to the zero position. In case periods are lost due to noise, the absolute position has to be traced back again to reset the counter. Several possibilities to create such an absolute position reference are applied in linear encoders. All options are based on an additional separate reference mark that is simultaneously exposed and can range from one simple set of



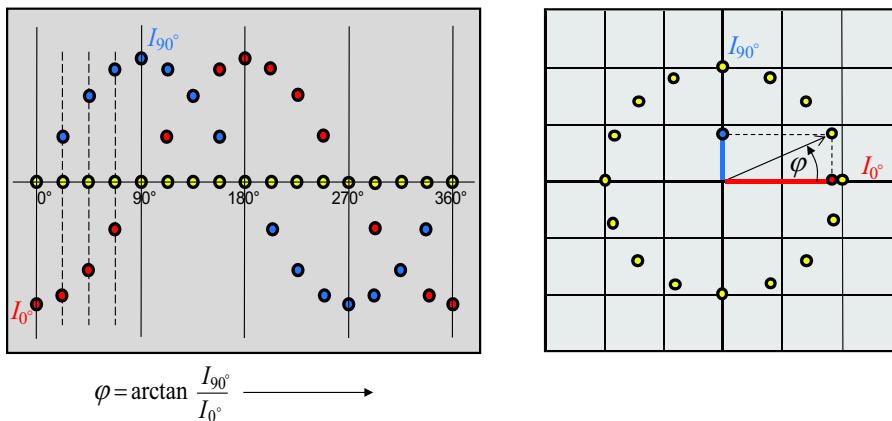
**Figure 8.65:** Sensor signals in the four field encoder as function of the relative position of the scales. At the left side the four positions of the measuring scale are shown relative to the scanning reticle, corresponding with the numbered vertical lines in the signal plot at the right side.

slits to a binary coded pattern that is read out by a CCD camera. As long as a transition of the reference mark uniquely points into a specific period of the incremental encoder signal, the absolute position can be deduced by setting the incremental counter accordingly.

The single field encoder has several problems that need to be addressed. First of all it is impossible to determine the direction of movement. Furthermore, the influence of electrical noise (dark current) and contamination of the gratings becomes large at the low signal level positions.

Figure 8.64 shows an enhanced version of this principle that solves these issues by the following measures:

- Redundancy by using two sections per phase for the same information but working in counter phase to have high signal levels at all positions.
- Two pairs of segments shifted 90° in spatial phase to give directional information.



**Figure 8.66:** Interpolation with two sine functions at  $90^\circ$  phase shift shown graphically by means of a Lissajous figure. The angle  $\varphi$  is proportional to the position within one cycle and can be calculated from both coordinate values.

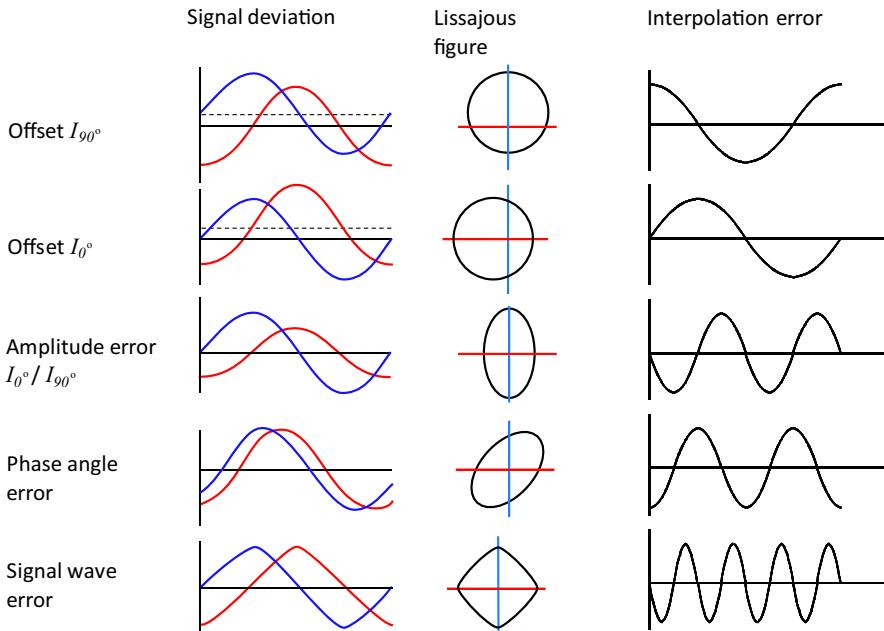
The beneficial effect of the redundancy is obvious. To better explain the effect of the phase shifted sections, Figure 8.65 shows the signal levels of the different fields, corresponding with the four positions where one of the fields is at its maximum irradiance level. It is shown that signal A will decrease while signal C has a positive sign when moving to the right from position 1 to position 2. A decrease of A with a negative sign of C (or more reliably a positive sign of D) corresponds with a movement from 3 to 2 which is in the left direction. This method gives a reliable direction information, but the total system can even be improved by a vectorial combination of all signals.

### 8.8.1.1 Interpolation

In principle the value of the position can be determined by digitising the current value of the field sensor by means of a comparator but then the resolution would not be better than one grating period.

The spatial sinusoidal character of the signal creates the possibility to make a reliable estimation of the intermediate positions. This estimation process is called interpolation and its principle is shown in Figure 8.66.

From the signals of the four field encoder two signals are created, one signal is the difference of A and B and the other is the difference from C and D. These signals can be plotted in a four quadrant vectorial representation, called a *Lissajous plot* after the French mathematician Jules Antoine



**Figure 8.67:** Deviations of the signals in their shape, magnitude and average value cause errors in the interpolation.

Lissajous (1822 – 1880). The phase angle  $\varphi$  appears to be equal to:

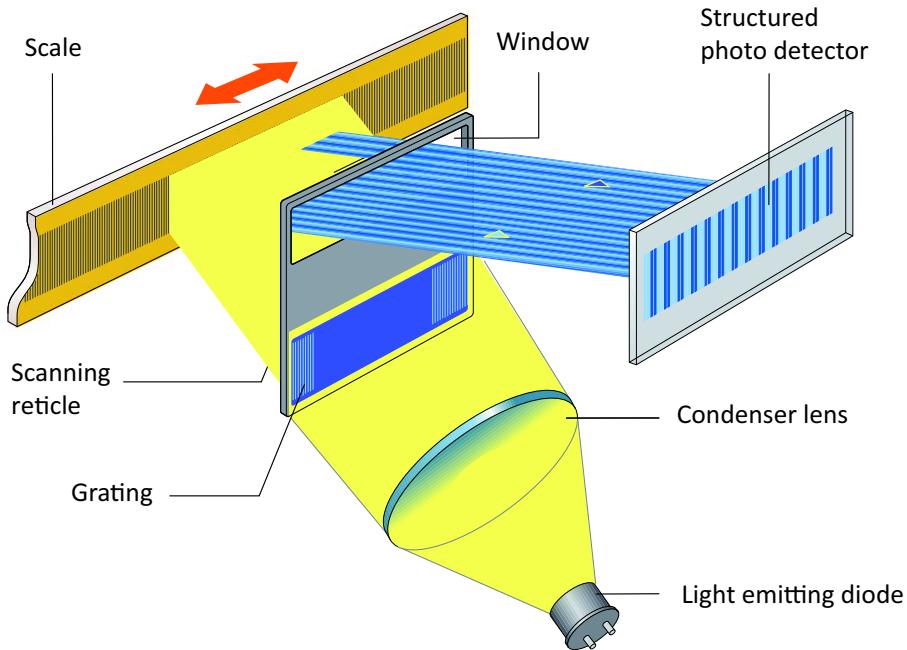
$$\varphi = \arctan \frac{I_{90^\circ}}{I_{0^\circ}} \quad (8.66)$$

where  $I_{0^\circ}$  equals the combined signal current of the sensor parts with their extreme values at  $0^\circ + n \cdot 180^\circ$  while  $I_{90^\circ}$  equals the combined signal current of the sensor parts with their extreme values at  $90^\circ + n \cdot 180^\circ$ .

Originally this calculation was done in the analogue domain by electronic addition of both signals with different amplification ratios, followed by a comparator. Nowadays the interpolation is far more easily achieved in the digital domain by using high resolution AD converters for digitising the signals followed by a direct calculation of the angle.

Under the condition of a fairly well defined sinusoidal shape of the signal an interpolation factor of 1024 until 4096 can be achieved.

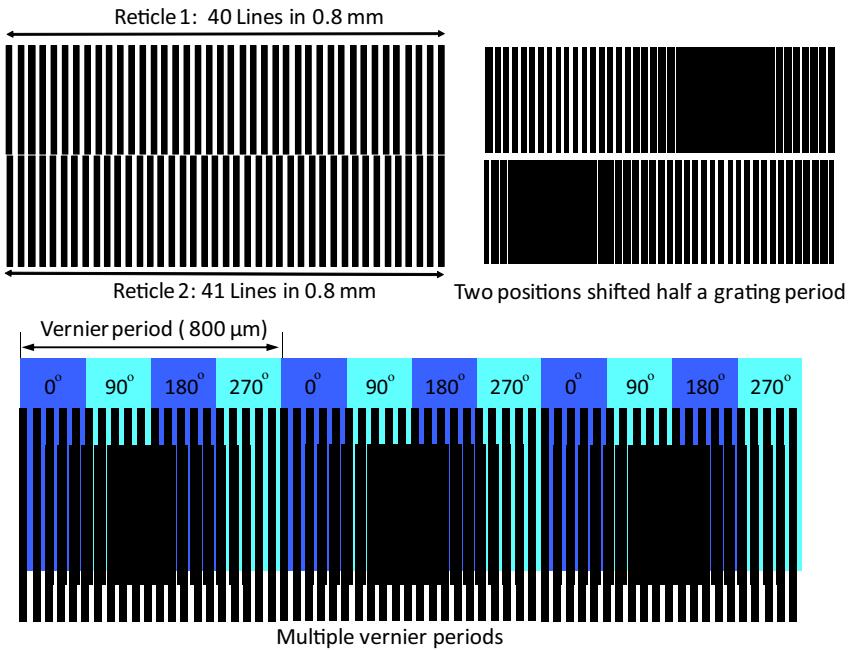
Figure 8.67 shows a qualitative overview of the errors that are caused by deviations from the ideal waveforms. A low interpolation error can only be achieved by means of careful mounting and positioning of all the optical parts in the measurement system.



**Figure 8.68:** The same properties can be obtained with a reflective scale as with a transparent scale. This alternative allows to use a very long flexible metal band scale that can be glued to the moving object. Directional information is obtained at the photo detector by a slight difference in periodicity of the gratings, causing a Vernier effect.

### 8.8.1.2 Vernier resolution enhancement

The previously described configurations have two important drawbacks. The first is the need for a transparent scale that requires mounting space of the sensor parts at both sides while complicating the connection with the moving object. The second drawback is the need for multiple sensor fields to realise both robustness and directional information. The first issue can be solved quite simply by using the reflective grating as shown in Figure 8.68. The configuration looks quite similar to the one field transparent grating and adds the benefit of the integration of the illumination and sensor part. This principle can be used with thin metal reflective scales that can be glued to the moving object. The inherent problem of the one field sensor regarding failing directional information and lack of ruggedness to contamination has been solved in a different way. By using two scales with a slightly different period a *Vernier* effect occurs, named after the French mathematician Pierre Vernier (1580 – 1637) who invented the principle. Like with a Vernier



**Figure 8.69:** Using Vernier gratings to obtain both directional and interpolation information with a one field CCD camera sensor.

calliper, this effect increases the position sensitivity and by reading out the signal with a structured photo detector like a CCD camera sensor, both directional and interpolation data are obtained.

The Vernier effect is illustrated in Figure 8.69. The upper half of the figure shows the effect of two gratings that differ one line over the total width of the grating of 40 lines. The corresponding Vernier ratio is 40 with a Vernier period that equals the shown size of the grating. The upper right image shows the visual effect when the two gratings are on top of each other in two situations that differ in a relative shift with half a grating period. It is clear that the dark and light zones have moved half the Vernier period giving a factor 40 more sensitivity over the movement of the grating. The moving direction of the dark and light zones is related to the relative moving direction of the gratings which enables to determine that direction from the signal of the CCD camera sensor.

Several Vernier regions can be combined to increase the robustness against contamination as shown in the lower half of Figure 8.69. As the entire surface of a sensor can only be used once, the designer can decide either for a very large Vernier ratio or for more Vernier regions for robustness. This

is an optimisation where the outcome depends on the application. With the resolution enhancement by the Vernier effect also an additional interpolation of the position can be done to a factor that depends on the resolution and noise of the applied CCD camera sensor.

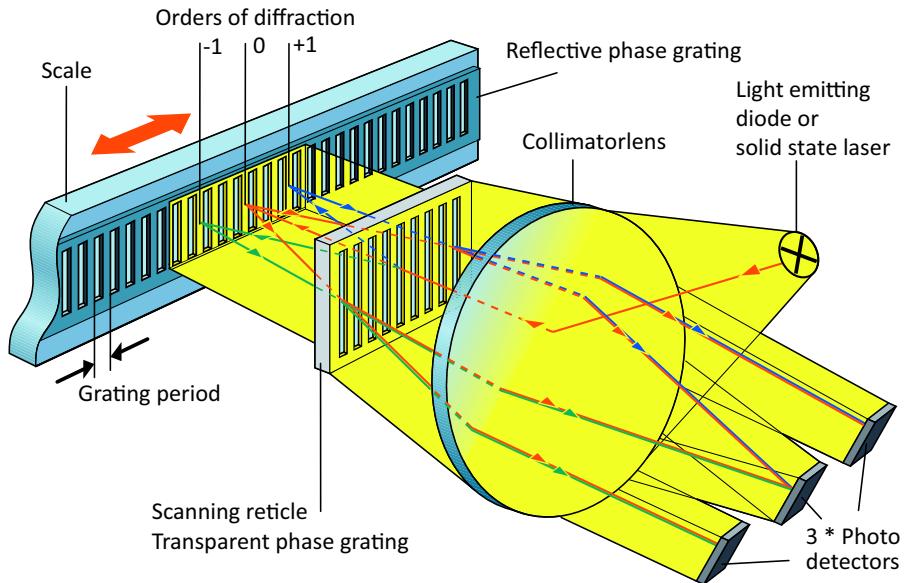
### 8.8.1.3 Interferometric optical encoder

The search for a continuous improvement in resolution, accuracy and robustness has resulted in manufacturing improvements to continuously decrease the grating period of the applied scales and eventually the grating period approached the wavelength of the light. As a consequence, simple shadow projection methods were no longer applicable and diffraction and interference effects had to determine the functionality.

The last example of a linear optical encoder is fully based on this effect of interference of light, using the phase gratings as described in Chapter 7 Section 7.4.3.2. This interferometric optical encoder enables resolutions into the sub-nanometre region and is applied in precision machines like wafer scanners.

The explanation of this more complex encoder principle will be done in three steps. First a global overview of the system is given. The second step explains the phase shift at the moving scale as function of the movement. In the third step all phase relations are determined that create the interference pattern on the photo detectors.

For the first step, Figure 8.70 is used to show the path of the light in three dimensions. The light source must be monochromatic with a high radiance to guarantee a clear phase relation of the light at the source. The coherence length does not need to be very long because the optical path length of all beams in the system is designed to remain equal. Small path differences may be induced by mechanical tolerances. Therefore the light source may have a rather short coherence length as long as these tolerances. A light emitting diode with a coherence length of  $\approx 15 \mu\text{m}$  is often sufficient but for more critical applications a solid state laser can be used. The diverging light from this point source is collimated into parallel rays by a positive lens and directed to the scanning reticle. The scanning reticle consists of a phase grating with a depth such that 75 % of the light is equally diffracted in three orders. The three orders are directed to the scale on the moving part with a reflective phase grating. The depth of this second phase grating is such that most of the light is diffracted into the 1<sup>st</sup>-order directions. This has different consequences for the diffracted orders from the scanning reticle. First of all the 0<sup>th</sup>-order from the scanning reticle will be reflected to a

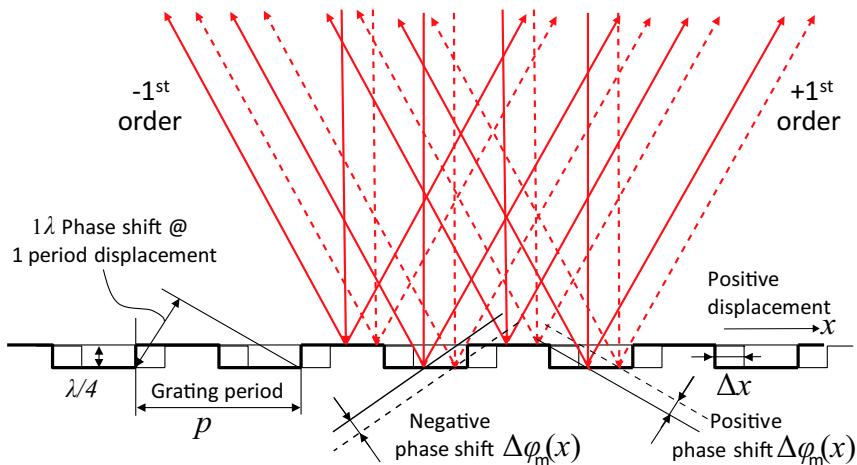


**Figure 8.70:** A phase grating with a periodicity in the order of the wavelength of the light result in a very high resolution and a three phase position signal because of the interference at the detectors.

+1<sup>st</sup>- and -1<sup>st</sup>-order at the scale. Secondly the ±1<sup>st</sup>-order directions from the scanning reticle will be reflected back orthogonal to the surface of the scale according to the fact that light, entering under an angle corresponding with the 1<sup>st</sup>-order, will be diffracted in the direction of the 0<sup>th</sup>-order.

Because of the symmetry and the equal diffraction angles, the resulting four beams are recombined at the scanning reticle where they will undergo a third diffraction with equal 0<sup>th</sup>- and ±1<sup>st</sup>-orders. Only two of these will be used further in the system. The third is not shown in the drawing and will in reality be projected by the collimator lens on an area without a sensor. The collimator lens that originally created the parallel rays from the light source will create one focused spot from all returning parallel rays on a position that is determined by the angle of these rays according to the principles of optical imaging. Depending on the angles the rays will come together at three photo detectors.

In the previous chapter on optics, Equation (7.26) showed that the radiance of the interference of two beams with an equal irradiance is proportional to  $1 + \cos \varphi$  with  $\varphi$  being the phase difference of the two beams. This means that the phase differences between all beams at the point of interference need to be determined.



**Figure 8.71:** A displacement of the phase grating causes a phase shift of the diffracted 1<sup>st</sup>- and -1<sup>st</sup>-orders with a sign equal to the sign of the order. One grating period displacement gives one wavelength phase shift of the diffracted orders.

For this reason in the second step of reasoning the effect of a displacement of the reflective phase grating on the phase of the reflected 1<sup>st</sup>- and -1<sup>st</sup>-order light is determined with the help of Figure 8.71.

During the displacement of the grating with one period the phase of the diffracted orders have also shifted exactly one period of the light. This means that the incremental phase shift in the refracted orders as function of an incremental displacement  $\Delta x$  is equal to:

$$\begin{aligned}\Delta\varphi_m(x) &= 360 \frac{\Delta x}{p} \quad [\text{deg}], \text{ or:} \\ &= 2\pi \frac{\Delta x}{p} \quad [\text{rad}],\end{aligned}\tag{8.67}$$

with a sign depending on the direction of movement relative to the direction of the order. With the indicated directions in Figure 8.71, the sign is equal to the sign of the order.

The third step uses the graphical representation of Figure 8.72 where a few simplifications are applied in respect to the real system to enable a more easy understanding of the total phase relationships between the different beams. First the collimator lens is left away. This is allowed as the collimator lens only creates the parallel beams and concentrates them again on the sensors. The phase impact of this lens, if any, is equal for all beams. Secondly the reflective grating is drawn as if it is a transparent grating. This is only for

graphical purposes as otherwise the reflected and incoming beams would overlap. The part right from the moving scale must be seen as mirrored<sup>9</sup> from the left and the second scanning reticle is in fact the same element as the first. The phase relations shown correspond with the real reflective grating. The third simplification is the omission of the phase shift due to the optical path length. Also this is allowed as the optical path length between the gratings is equal for all beams. The last simplification is the omission of the diffraction orders that are not used. This has some consequences as it might seem that not all orders will keep the same radiance level. Where necessary the effect is separately mentioned.

Starting at the left side, first the incoming light from the light source is diffracted by the scanning reticle into three orders with an equal irradiance level of 25 % each. The remainder of the light is emitted in higher orders. It was shown in Section 7.4.3.2 of Chapter 7 that such a phase grating will cause a phase shift of  $+60^\circ$  at both 1<sup>st</sup>-orders and a phase shift of  $-60^\circ$  for the 0<sup>th</sup>-order. Like presented before, the moving scale is a 50 % diffracting scale that does not introduce a phase shift by the diffraction but only by the incremental movement  $\Delta x$ . This incremental phase shift is  $-\Delta\varphi_m(x)$  for a diffraction angle opposite to the motion direction and  $+\Delta\varphi_m(x)$  for a diffraction angle in the motion direction. As shown in step one, the original 1<sup>st</sup>-orders from the first transition through the scanning reticle will be diffracted by the moving scale into the 0<sup>th</sup>-order direction orthogonal to the scale and another order that is not used. The orthogonal order is recombined at the scanning reticle with both 1<sup>st</sup>-orders that originated from the 0<sup>th</sup>-order of the first transition at the scanning reticle.

The real important phase shifts happen at the second transition through the scanning reticle. All beams will be diffracted in three orders of which only two are used. Depending on the diffraction angle there will be an additional  $+60^\circ$  phase shift when diffracted under an angle and a  $-60^\circ$  phase shift when running in the same direction as the incoming beam. After the second transition of the scanning reticle four beams are obtained that each consist of two beams with a different phase, each determined by the addition of all phase shifts over their respective optical path. The phase difference between both beams in a pair gives an irradiance value at the photo detectors with a cosine function of double the phase shift by the moving scale. This doubling is a useful increase of the sensitivity. The three irradiance signals of the detectors have a  $120^\circ$  phase relation with each other because of the different  $60^\circ$  phase shifts. This three phase measurement signal is

<sup>9</sup>This method for graphical presentation of an optical system is often applied in situations where mirrors are applied.

even more optimal for interpolation, than the 90° shifted signals of the four field encoder. There is even less need for inverse sensors to overcome the noise at the low irradiance areas because at any position sufficient signal with a high irradiance level is available to increase the signal to noise ratio. There is also no need for complicated structured sensors with Vernier enhancement. Furthermore the sensitivity for contamination is acceptable, as contamination will only cause a reduction of the irradiance by scattering of light but the phase relationship is not affected. The irradiance levels of the different beams will however influence the interpolation and a very strict adherence to narrow tolerances is required when the interferometric encoder is used at extreme levels of interpolation.

#### **8.8.1.4 Concluding remarks on linear encoders**

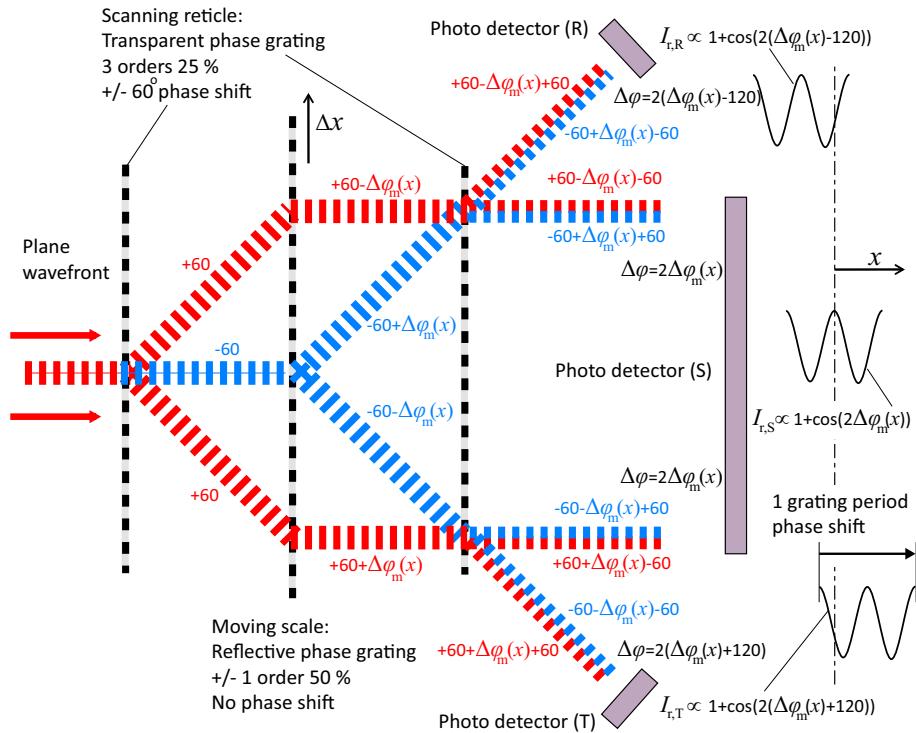
Only a limited overview is presented of the numerous methods to achieve incremental position information by means of a grating scale with dedicated sensing optics.

An interesting method that for instance not has been mentioned is the two dimensional sensing principle of the optical computer mouse. This inexpensive system uses a CCD sensor to determine the displacement of a surface by means of its surface structure. This structure fulfils the role of grating and the pattern can be of any shape. Improvements with a laser source to enhance contrast have made these sensors quite reliable and sensitive up to sub-micrometre levels. Unfortunately this method is very sensitive for distance variations, which could be solved with telecentric optics but that will dramatically increase the cost. Nonetheless, it surely is an option for future sensors in more advanced applications.

An important benefit of optical encoders when used in high precision positioning systems is the small distance between the sensor and the scale. This largely reduces the sensitivity for changes in the index of refraction of the intermediate air. This problem is more prominently present in the laser interferometers that are presented in the following section.

With a further enhancement of the interferometric encoder principle it is even possible to measure in more directions with one integrated sensor and a two dimensional grating. In Chapter 9 such a sensor is used to measure the position of the wafer stage in a wafer scanner.

Optical encoder measurement systems can be applied to measure very large displacements with a high resolution as long as the scale can be mounted on the moving object and the sensor can keep track of the scale. This limitation is not present with measurement systems that are based on the determina-



**Figure 8.72:** Graphical representation of the phase of an interferometric optical encoder. The phase shift  $\varphi_m$  caused by the moving scale gives an irradiance pattern  $I_r$  at the photo detectors that gives a signal of two spatial periods for one period movement of the scale. The  $60^\circ$  phase shifts at the transparent grating create  $120^\circ$  spatial phase difference between the three sensors, resulting in a three phase measurement signal (R,S,T).

tion of the optical path length differences by means of laser interferometers.

### 8.8.2 Laser interferometer measurement systems

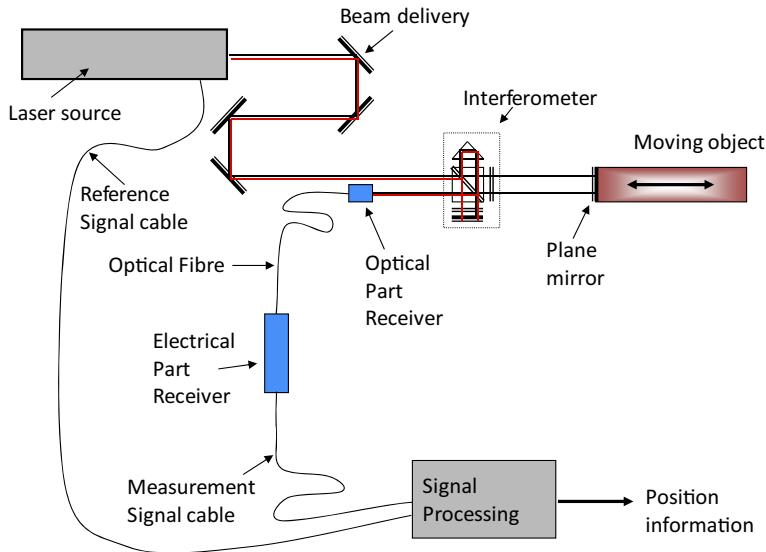
In the previous section it was demonstrated that incremental information on the position can be obtained by interference. In laser interferometry the interference is not caused by combining diffracted beams of light but by combining beams of light that passed through optical paths with a different length. Laser interferometry is an extremely wide field of expertise with countless configurations. Two quite different application areas can be distinguished for geometric measurement.

First of all laser interferometry is the “de facto” standard for the measurement of the surface shape topology of optical elements and precisely machined objects. By comparison of the phase of a wavefront of light that is reflected from the measured surface with the phase of the wavefront that is reflected from a known surface, deviations of up to several picometres can be determined. These measurements are in principle static and require extreme precautions for vibrations and stability.

For mechatronic systems that deal with controlled motion, the second application area that focuses on displacement measurement of solid objects is more important. For this reason this section presents only displacement measurements of moving objects with even only the most frequently used principle, the interferometer based on the Michelson configuration. Though only one of many possible configurations, its understanding gives sufficient background knowledge to understand also the physical principles of the other interferometers.

Figure 8.73 shows an overview of the main components of a laser interferometer displacement measurement system<sup>10</sup>. A laser source is used with a large coherence length, because of the interference principle. The phase relationship of the light needs to remain coherent over the entire optical path which is far longer than with the interferometric encoder of the previous section. The real measurement of the distance to a moving object takes place at the interferometer itself. The source of light is split in the interferometer in two beams with a known phase and frequency relationship. One beam travels to a reference reflector and the other to a reflector to the object of which the distance is to be measured. After returning to the interferometer both beams are mixed and interfere. The resulting interference pattern is an irradiance-modulated light signal as function of the phase difference between the interfering beams. This interference signal is less critical in

<sup>10</sup>Several of the used figures and background information in this section are graciously provided by **Agilent Technologies**, a leading company in extreme precision multi-axis laser interferometer displacement measurement systems.



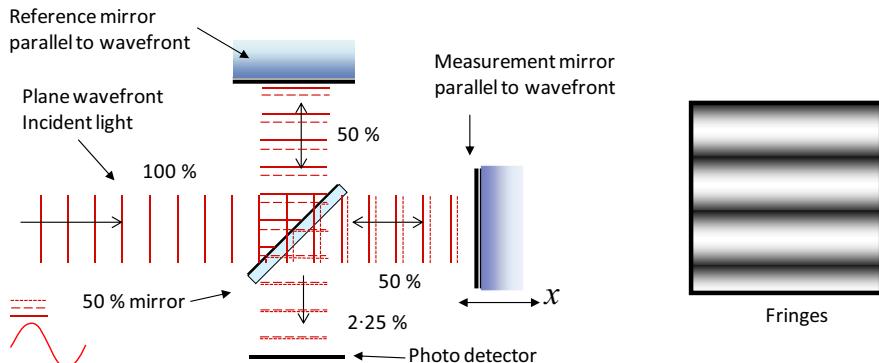
**Figure 8.73:** The main components of a laser interferometer measurement system. The reference signal cable is only used with the heterodyne principle.

respect to wavefront errors so it can be transported to a photo detector via an optical fibre preceded by an optical part that focuses the light into the fibre. The fibre is connected to the receiver electronics that converts the irradiance of the light into an electrical signal and sends it to the signal processing element.

### 8.8.2.1 Homodyne distance interferometry

The most simple version, the *homodyne interferometer* is shown in Figure 8.74 where the term “homodyne” refers to the use of light with only one frequency. The monochromatic coherent light of a laser enters the system from the left and is split in two parts with equal irradiance by a 50 % reflecting mirror. The first part, the reflected light, goes towards the reference mirror, where it is reflected back to the 50 % mirror that will again split the light in two equal amounts. One half of the light is reflected back to the source and not used anymore. The remaining light with an irradiance level of 25 % of the level at the input of the system, will pass the mirror and reach the photo detector.

The second part, the transmitted light from the first encounter with the 50 % mirror, goes towards the measurement mirror and likewise with the first part, this beam will be reflected back to the 50 % mirror and again half



**Figure 8.74:** Basic homodyne Michelson Interferometer with a 50 % reflecting mirror, a reference mirror and a measurement mirror. The reflected light after the reference and measurement mirror is dashed to show the paths. At the right, the irradiance pattern is shown with fringes that occur at the photo detector plane when the wavefronts are not running parallel. Note the phase jumps of  $0.5\lambda$  at the mirrors when reflecting from low refractive index to high refractive index material.

of the light is lost and transmitted back to the source with only 25 % of the light left to be reflected at the mirror and reach the photo detector. When all optics are well aligned, the wavefronts of both beams are still parallel and interference takes place on the photo detector, either constructive when both beams are in phase or destructive when both beams are out of phase or partial in the situation in between. The reference mirror is stationary so only the phase of the reflected beam from the measurement mirror is modulated by the distance. A displacement results in a sinusoidal irradiance modulation of the interference at the photo detector according to the  $1 + \cos\varphi$  relation from Equation (7.26) in the previous chapter on interference. Because of the double-passing of the trajectory from the interferometer to the measurement mirror and back the incremental phase shift related to an incremental movement  $\Delta x$  of the measurement mirror equals:

$$\Delta\varphi_m(x) = \frac{2\pi N}{\lambda} \Delta x = \frac{2\pi f_p N n}{c} \Delta x \quad [\text{rad}], \quad (8.68)$$

where  $N$  equals the *interferometer constant* defined as the number of single trajectories that the measurement beam passes to and from the object, in this case  $N = 2$ .  $\lambda$ ,  $f_p$  and  $c$  are respectively the wavelength, the frequency and the propagation velocity of the light in vacuum while  $n$  equals the refractive index.

During a movement, the irradiance at the photo detector equals zero in

every period of the modulated interference. Apparently the energy from the source has disappeared at those positions. In reality the law of conservation of energy abides. It appears that the light of both beams that is sent back to the source shows constructive interference when the light at the detector shows destructive interference and the other way around. This can be checked when realising that reflection at a surface from a low refractive index medium to a high refractive index material introduces a sign reversal of the phase. This happens at the reference mirror, the measurement mirror **but only at one side of the 50 % reflecting mirror**, the side on the mirror glass with the reflective aluminium coating. When counting the phase sign reversals at the mirrors, tracing both beams, it shows that the beam at the photo detector that came from the reference mirror has undergone two sign reversals while the beam from the measurement mirror had undergone one sign reversal. This means that the sign of the phase difference between both beams due to the displacement is reversed. The beams that are reflected back to the source have either one sign reversal from the measurement mirror while the beam from the reference mirror got a total of 1.5 sign reversals. This means that both beams have the same sign reversal and as a consequence the sign of the phase **difference** between both beams due to the displacement is not reversed. This means that the phase difference in the beams to the sensor and to the source differ exactly one sign reversal. As a consequence, constructive interference in one of the two beams corresponds with destructive interference in the other beam.

### Directional information and interpolation

The described basic homodyne interferometer is comparable with the basic optical encoder with only two gratings and one sensor. At standstill the signal from the detector is a constant value and with a movement a sinusoidal frequency is present that is proportional to the velocity. Unfortunately, like with the basic encoder, there is no directional information and also at low irradiance levels the noise impairs the accuracy.

One method to solve these problems is comparable with the previously presented Vernier encoder method with a structured sensor. One could say that the Vernier effect is caused by an imperfection in the periodicity of the gratings. Along the same line of thoughts the solution in the interferometer is based on an imperfection in the orientation of the mirrors.

A perfect constructive or destructive interference happens only when the wavefronts are perfectly flat and aligned. As soon as one of the mirrors is slightly tilted, *fringes* are observed at the photo detector as shown at the

right side of Figure 8.74. These are caused by differences in path length due to the tilt<sup>11</sup> and their spatial frequency is proportional to the tilt angle. With a single photo detector these fringes would result in a constant average irradiance over the surface of the sensor when the fringe period is smaller than the diameter of the sensor. In a normal interferometer this would require a smaller sensor at the sacrifice of sensitivity but it is far better to use a structured sensor like a CCD camera. In that case the movement direction can be detected as these fringes will move over the surface of the detector when the measurement mirror moves in the  $x$ -direction. A displacement of  $\Delta x = \lambda/N$  will cause a shift of the fringes with one fringe period. This enables a good resolution with sufficient interpolation capability of the spatial sinusoidal signal. An interpolation factor of hundred already yields a resolution of 3 nm with a Helium-Neon laser of 632 nm.

Although with further improvements on interpolation even better results are obtainable, this principle still has some additional drawbacks that have challenged people into different concepts to solve these. First of all there is the sensitivity for the tilt in the mirrors and secondly the signal frequency ranges from 0 Hz until very high frequencies depending on the movement velocity.

## Sensitivity for tilt

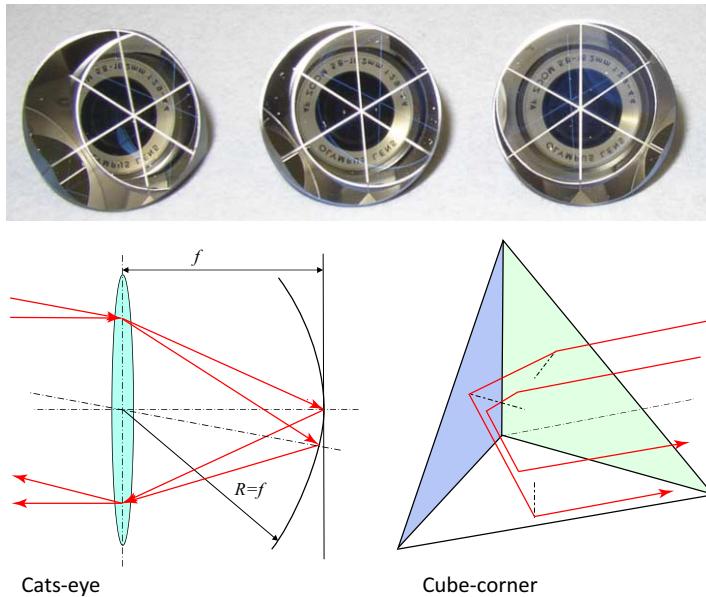
Sensitivity for the tilt of the reference mirror can be avoided by a very stable mounting of this mirror in the interferometer. Tilt of the measurement mirror will, however, directly influence the observed effect on the CCD sensor. This would limit its use to systems with only one degree of freedom on a perfect straight guiding mechanism that does not introduce angular movements.

This problem can be avoided by replacing the measurement mirror by a *cube-corner retro-reflector*, also called just a cube-corner or corner-cube. Also the *cats-eye* can serve this purpose. Both optical systems have in common that they reflect the light back (retro) in the same direction as where it came from without any optical path difference for parallel rays, thus keeping the reflected wavefront flat and parallel to the incident light. Figure 8.75 shows both optical systems.

A cats-eye consists of a positive lens with a spherical mirror in its focal point. The radius of the spherical mirror is equal to the focal length of the mirror

---

<sup>11</sup>This effect of these fringes is primarily used when measuring surfaces with interferometers. The fringe shape and distance represents the value of the surface slope relative to the reference.



**Figure 8.75:** A cats-eye and a cube-corner retro-reflector. A cats-eye consists in its basic shape of a positive lens with a spherical mirror in the focal point of the lens. The cube-corner is a prism with three mutually orthogonal reflecting surfaces. With both optical systems, the light rays will be reflected in the same direction as where they came from but shifted to an opposite position of the chief ray through the central point of the optic entry pupil. This point symmetry is shown in the photographic picture from three cube-corners with a different orientation where the reflection of the lens always has the same orientation to the camera. The text on the lens, however, is mirrored.

causing the chief ray through the centre of the lens to always be reflected back in the same direction. As the reflecting surface is positioned in the focal point of the lens all parallel rays will be focused on one point on the spherical surface and reflected back under the same angle with the chief ray which is always orthogonal to the spherical surface.

After the lens the reflected rays will be re-collimated parallel to the incident rays. Their position is however mirrored in respect to the point of incidence of the chief ray. For this reason this system is called *point-symmetric*. The main drawback of a cats-eye is the fact that a simple two element cats-eye can only be used with small apertures and approximately paraxial rays as otherwise spherical aberration and coma will cause wavefront errors. With aspheric optics or multiple optical elements these errors can be avoided at

increased cost.

The cube-corner has less problems in that respect because it does not apply curved surfaces. It consists only of a monolithic piece of transparent material with three mutually perpendicular mirror surfaces that internally reflect all rays back in the same direction. Like the cats-eye, the cube-corner is a point-symmetric retro-reflector.

### Wide frequency range

The wide frequency range of the interference signal is typically problematic for low signal level situations as noise can more easily be filtered out when a limited frequency range is used. This low signal-level situation is typical the case in a multi-axis measurement system where one laser source has to supply a multitude of interferometers. Encoders often consist of a complete system per direction, including the light source and sensor, and in that case the noise level is less a problem.

It is especially important to avoid a DC value of the signal as all DC measurements show some level of DC drift over time. In Section 8.4.2 it was shown that by modulating a low-frequency signal, the spectrum of the measurement signal can be shifted to a higher frequency. With a homodyne interferometer this modulation can be realised by adding a small high-frequency motion oscillation to the reference mirror, resulting in a modulated signal at the detector. This can be synchronously demodulated after amplification.

It is also possible to use polarisers and polarisation dependent optical elements to create two signals with  $90^\circ$  phase difference to get the directional information by quadrature detection, a method that enables a plotting of the signal as a vector in a four-quadrant plane like the Lissajous plot of Figure 8.66 that was introduced with the theory on interpolation of encoder signals.

Ultimately, *heterodyne interferometry* appeared to be a better solution for high precision displacement measurements with laser interferometry in actively controlled positioning systems.

#### 8.8.2.2 Heterodyne distance interferometry

The heterodyne distance interferometer solves the above mentioned problems by using a laser source with two different beams, one for the reference and one for the measurement path that have a different frequency and a mutually orthogonal linear polarisation direction. In most cases the beams are combined into one beam at the input of the system but the principle can

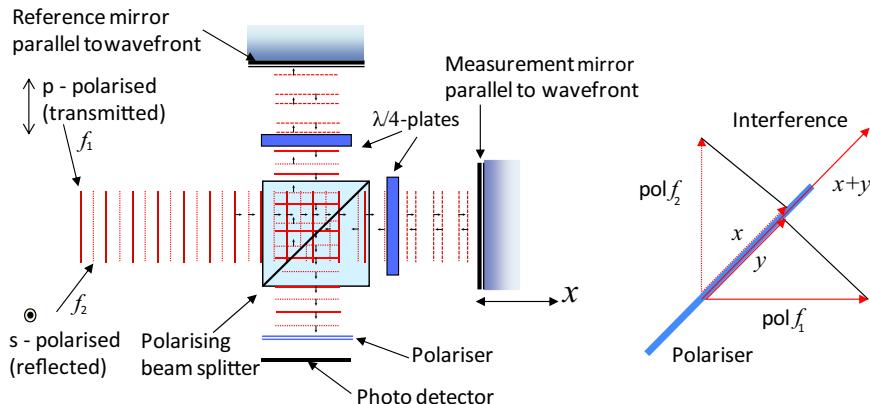
also work with fully separated beams as long as they have a known phase and frequency relationship.

The term “heterodyne” refers to the frequency difference. This *split-frequency*  $f_s$  can be in the order of several Megahertz. One method to generate a laser source with such a dual frequency spectrum is based on the Zeeman effect, named after the Dutch physicist and Nobel prize winner Pieter Zeeman (1865 – 1943) who discovered the effect. When the laser cavity of a Helium Neon laser is inserted in a permanent magnetic field, the normally single laser-frequency radiation is split in two radiation parts with an equal irradiance but with two different frequencies and polarisation states. The two frequencies are symmetrically spaced around the original frequency, called the *centre-frequency*  $f_c$ . The split-frequency is proportional to the magnetic flux density. The polarisation direction of both radiation components is circular with opposite directions. This circular polarised light is converted into two orthogonal linear-polarised radiation components by a birefringent plate with one quarter of a wavelength delay between the orthogonal polarisation directions. This *quarter wave plate* or  $\lambda/4$ -plate plays an important role in the further explanation of the heterodyne interferometer.

A disadvantage of the Zeeman method to create a laser beam with two frequencies is the limited optical power. This is due to the bandwidth of the Fabry-Perot cavity from the laser, which is too small to optimally resonate at both frequencies. This effect is especially prominent when a very high split frequency is needed. As will be shown later, measuring displacements with a very high velocity demands such a high frequency.

In case more power is needed at such a high split-frequency, a suitable light source can be created by modulation of the light of a single-frequency laser by means of an acousto-optic modulator. An acousto-optic modulator contains a piezoelectric actuator that creates running sound waves in a material like quartz. The resulting density differences act like a modulated moving refractive-index phase grating. Depending on the intensity of the modulation, the magnitude of the light in the diffraction orders is modulated in a similar way as was presented on modulation by the depth of a phase grating. More important is the frequency shift that is caused by a Doppler effect at the moving sound waves that creates different frequencies for the different diffraction orders. This frequency difference equals the frequency of the sound times the order number.

Figure 8.76 shows the principle of the heterodyne interferometer. The different polarisation directions of the two-frequency beam components are used to guide the light to the measurement and reference optical path in the following way:

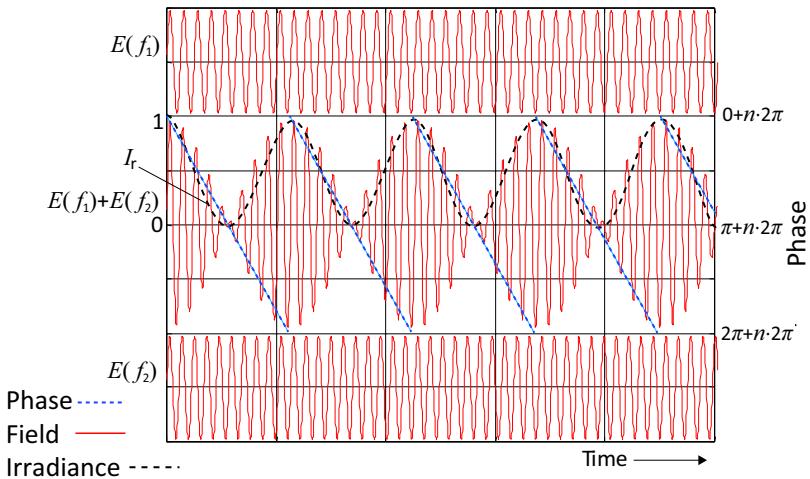


**Figure 8.76:** Heterodyne Michelson interferometer with a two frequency orthogonal polarised laser where the polarisation is used to separate the reference and measurement path by means of a polarising beam splitter and two  $\lambda/4$  plates. At the right side the effect of the polariser is shown enabling interference that otherwise would not occur between orthogonal polarised beams.

Note that the phase steps of  $180^\circ$  at the reflecting surfaces are not shown as this is not relevant for this principle.

The combined beam enters the system from the left. The beam component with frequency  $f_1$  is p-polarised in the plane of incidence of the *polarising beam splitter* and the beam component with frequency  $f_2$  is s-polarised. S-polarised light will be reflected by the polarising beam splitter and p-polarised light will be transmitted. As a consequence the s-polarised light will serve as reference beam. It is reflected towards the reference mirror and passes a quarter wave plate consisting of a birefringent material. When oriented in the right way to the polarisation direction of the beam, this  $\lambda/4$  plate creates a phase lag of  $90^\circ$  in one of the polarisation directions relative to the other direction as explained in Section 7.4.1.1 in the previous chapter and changes the polarisation from linear into circular. After reflection at the reference mirror the second pass through the  $\lambda/4$  plate will convert the circular polarised light again into linear polarised light but orthogonal to the original direction. As a consequence the now p-polarised beam will be transmitted through the polarising beam splitter towards the polariser and the photo detector.

The p-polarised component of the light from the laser source serves as measurement beam. It is first transmitted by the polarising beam splitter and after a passing through a second  $\lambda/4$  plate, reflection at the measurement



**Figure 8.77:** Two interfering beams with different frequencies result in an irradiance signal that is proportional to one plus the cosine of  $2\pi(f_1 - f_2)$ . The phase between both frequencies shifts linear over time.

mirror and a second pass through the  $\lambda/4$  plate it has become s-polarised and is reflected towards the polariser and the photo detector where it recombines with the reference beam.

The polariser fulfils an important requirement as the vectorial addition of orthogonal polarised electromagnetic fields will not show destructive or constructive interference. With an accurately oriented polariser under  $45^\circ$  with both polarisation directions only the component of each beam in the polarisation direction of the polariser will pass as shown in the right drawing on Figure 8.76. This means that after the polariser the fields of both beams have an equal direction and the vectorial addition will show interference at the expense of a factor  $\sqrt{2}$  loss of amplitude which corresponds with a factor two of the radiance. It can be reasoned that also in this case constructive interference at one location corresponds with destructive interference at another place, hence corresponding to the physical law of conservation of energy. In case  $x$  and  $y$  are in phase in the direction of the polariser as drawn, the components of  $\text{pol}f_1$  and  $\text{pol}f_2$  in the orthogonal direction are in counter phase.

## Interference at the detector

The resulting interference field at the detector consists of the two fields with different frequencies, like shown in Figure 8.77, that add according to the following<sup>12</sup> equation:

$$E_t = E_1 + E_2 = \hat{E}(\sin(2\pi f_1 t) + \sin(2\pi f_2 t)) \quad (8.69)$$

where  $\hat{E}$  is the equal field amplitude of both fields.

Using a trigonometric identity and cancellation of the factors two in the numerator and denominator yields the following equation:

$$\begin{aligned} E_t &= 2\hat{E} \sin\left(\frac{2\pi(f_1 + f_2)t}{2}\right) \cos\left(\frac{2\pi(f_1 - f_2)t}{2}\right) \\ &= 2\hat{E} \sin(\pi(f_1 + f_2)t) \cos(\pi(f_1 - f_2)t) = 2\hat{E} \sin(2\pi f_c t) \cos(\pi f_s t), \end{aligned} \quad (8.70)$$

From this equation the sine term is the very high-frequent centre frequency. The cosine term is the much smaller frequency difference and can be seen as the amplitude modulation of the high-frequency term.

The irradiance is proportional<sup>13</sup> to the field magnitude squared. Because the high-frequency term is unmeasurably high it becomes a constant RMS value of  $\sqrt{2}$  that can be left out in the proportional equation, resulting in the following expression for the irradiance:

$$\begin{aligned} I_r &\propto \hat{E}^2 \cos^2(\pi f_s t) \\ &\propto \hat{E}^2 \left(\frac{1 + \cos(2\pi f_s t)}{2}\right) \propto \hat{E}^2 (1 + \cos(2\pi f_s t)) \end{aligned} \quad (8.71)$$

The resulting signal after the photo detector is a combination of a DC voltage and a sinusoidal voltage with a frequency that is equal to the frequency difference of the two beam components that interfered at the sensor.

When comparing this equation with the  $1 + \cos\varphi$  relation from Equation (7.26) that is used with the homodyne interferometer it is clear that the two heterodyne frequency components represent a continuous phase shift:

$$\varphi_s = 2\pi f_s t \quad (8.72)$$

The phase shift is linear proportional with time at standstill.

It will be shown in the following that this phase relation is changes by a

---

<sup>12</sup>The following mathematical analysis including the lock-in quadrature detection is derived from the Phd thesis of Jonathan Ellis from our laboratory in Delft. His thesis can be downloaded for more details.

<sup>13</sup>The constant  $cne_0$  is avoided in these expressions as it does not contribute to the understanding of the principle.

change in the optical path difference between the measurement and reference beam which enables the measurement of movements with only the AC part of the signal, thereby avoiding induced errors by low-frequency noise.

### Velocity and position detection by Doppler shift

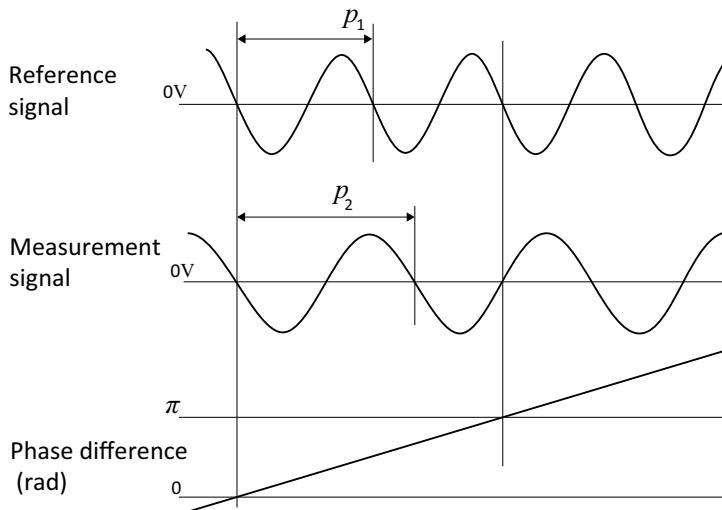
When the measuring mirror in a heterodyne interferometer moves along the direction of the beams, the measured frequency difference after interference  $f_m$  becomes different from the frequency difference  $f_s$  at the input of the interferometer. This phenomenon is called the *Doppler shift* of the measurement beam, caused by the Doppler effect that was discovered in 1842 by the Austrian physicist Christian Andreas Doppler (1803 – 1853). The Doppler effect is related to the constant wavelength and propagation velocity of a travelling wave. At a fixed position a travelling wave with propagation velocity  $v_p$  and wavelength  $\lambda$  will be observed as a fixed temporal frequency  $f = v_p/\lambda$ .

When the position of the observer is changing, the observed frequency is decreased when moving in the same direction as the wave propagation and increased when moving in the opposite direction. In fact the spatial frequency of a wave, as determined by the wavelength, is converted into a temporal frequency by the movement of the observer. Depending on the movement direction, this temporal frequency is added to or subtracted from the temporal frequency of the wave that would be observed at stand still.

A light wave with temporal frequency  $f_c$  and a propagation velocity  $v_p = c/n$ , with  $n$  being the refractive index, experiences the Doppler shift in the frequency of the measurement beam by a motion velocity  $v_m$  equal to the following expression:

$$\begin{aligned} f_d &= f_c \left( 1 - \frac{v_p}{v_p \pm Nv_m} \right) = f_c \left( 1 - \frac{c/n}{c/n \pm Nv_m} \right) \\ &= f_c \left( \frac{c/n \pm Nv_m - c/n}{c/n \pm Nv_m} \right) \approx \pm f_c \left( \frac{Nnv_m}{c} \right) \end{aligned} \quad (8.73)$$

The approximation is valid when the motion velocity is very small in respect to the propagation velocity of the wave which is in practice true for the speed of light.  $N$  equals the previously defined interferometer constant and presents itself as a multiplication factor for the Doppler effect, corresponding with the number of times that the measurement beam travels the path to the moving mirror. Furthermore the centre frequency  $f_c$  is in most cases allowed to be taken as reference as the split frequency is with even 5 MHz only a small part  $\approx 10^{-8}$  of  $f_c$ . If necessary this small deterministic factor



**Figure 8.78:** The velocity and position of the measurement mirror can be detected by measuring the difference in the timing period of the reference and measurement irradiation signal or by measuring the relative phase.

can be taken into account. More important is the refractive index  $n$  that shows to be a factor that directly influences the observed frequency shift. It will be shown later that the refractive index induces a serious measurement uncertainty when measuring in air with extreme requirements on precision. The Doppler shift is equally present as an offset in the frequency difference at the measurement sensor  $f_m = f_s + f_d$ . This means that  $f_d$  can be detected by comparing the detector signal from the interferometer with a reference signal equal to the split frequency  $f_s$  of the beam components at the entry of the interferometer. With a Zeeman laser this difference signal is obtained by creating an interference signal from a small portion of the light via a beam splitter, a polariser and a detector, similar to the detection after the interferometer. With an acousto-optic modulated laser the frequency of the modulator can directly be used as reference.

The detection can be done in several ways. One method uses a comparator like a Schmitt trigger to detect the zero voltage crossings of the AC part of the detector signal and the reference signal. When the timing period between the zero crossings of both signals is precisely measured, as indicated in Figure 8.78, the difference between both signal periods is an accurate measure for the frequencies and the corresponding velocity.

When this measurement is done at regular intervals, the incremental position change over that interval is determined by multiplying the interval

period with the found velocity.

In principle the sampling of the zero crossings is equal to a phase measurement as with different frequencies their phase relation shifts proportional with the time as function of the velocity. When applying Equation (8.72) for the measured frequency  $f_m$ , the following relation for the phase is obtained:

$$\varphi_t = 2\pi f_m t = 2\pi(f_s + f_d)t = \varphi_s + \varphi_d \quad (8.74)$$

The phase term  $\varphi_s = 2\pi f_s t$  is equally present in the reference signal which means that, when the measurement and reference frequencies are compared in phase, only the Doppler phase shift  $\varphi_d$  remains. With Equation (8.73) and neglecting the small error due to the velocity difference between the movement and the speed of light this becomes:

$$\varphi_d = 2\pi f_d t = \pm 2\pi f_c \left( \frac{N n v_m t}{c} \right) \quad (8.75)$$

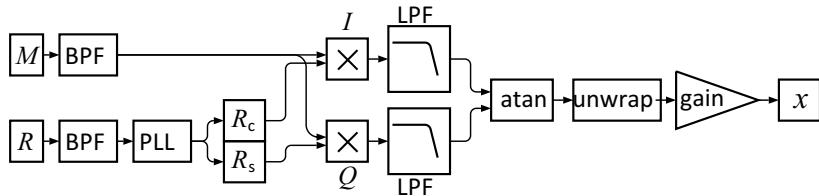
Within an incremental time period  $\Delta t$  the incremental displacement equals  $\Delta x = v_m \Delta t$  and the incremental phase shift due to this displacement becomes:

$$\Delta\varphi_m(x) = \pm \frac{2\pi f_c N n}{c} \Delta x \quad (8.76)$$

This relation is equal to the phase relation that was found with the homodyne interferometer. This is not without logic as the heterodyne frequency difference only introduces an additional phase shift to the phase shift caused by the displacement. For this reason the second detection method measures directly the incremental phase shift as function of an incremental displacement by means of *quadrature detection* with a lock-in amplifier.

### Lock-in quadrature detection

Figure 8.79 shows a schematic of a typical lock-in amplifier used to measure the phase between two signals. Both the measurement ( $M$ ) and reference ( $R$ ) signals are bandpass filtered (BPF) around the frequency  $f_s$  to remove their nominal offset, eliminate the sensitivity to optical power fluctuations, and provide anti-aliasing filtering for the digital signal processing. After filtering, the reference signal is sent to a phase-locked loop (PLL). A phase-locked loop is a feedback controlled variable oscillator of which the frequency is controlled with a phase detector to track the input frequency with a next-to-zero phase difference. The reason to apply a phase-locked loop is



**Figure 8.79:** Schematic of a lock-in amplifier for phase measurements. Two signals are detected and initially filtered. The reference signal  $R$  is sent to a phase-locked loop (PLL) to generate matched sine and cosine signals. Those are then multiplied with the filtered measurement signal  $M$  to produce in-phase ( $I$ ) and quadrature ( $Q$ ) outputs. Those are then low-pass filtered and sent to an arctangent function. The phase is then unwrapped and a gain is applied to determine the displacement  $x$  from the unwrapped phase.

(Courtesy of Jonathan Ellis)

its possibility to generate two signals with a nominal phase offset of  $90^\circ$ . Similar to the Lissajous plot shown with the spatial signals of a decoder with four sensors, two temporal signals with  $90^\circ$  phase difference enable signal interpolation and directional information. The name quadrature detection for this method comes from radio transmission practice and is related to the four quadrant vectorial representation defined by the phase difference of the signals.

The two outputs with  $90^\circ$  phase difference from the phase-locked loop,  $R_c$  and  $R_s$ , are multiplied with the measurement signal to generate two signals, the in-phase  $I$  and the quadrature  $Q$  signals. These signals relate in the following way to a phase shift  $\varphi_m(x)$  related to the difference in optical path lengths between the reference and measurement beams:

$$I = R_c M = \cos(2\pi f_s t) \cos(2\pi f_s t + \varphi_m(x)) \quad \text{and} \quad (8.77)$$

$$Q = R_s M = \cos\left(2\pi f_s t + \frac{\pi}{2}\right) \cos(2\pi f_s t + \varphi_m(x)), \quad (8.78)$$

where  $\varphi_m(x)$  is the Doppler phase change of the measurement signal, relative to the reference signal as a function of the displacement. Applying a trigonometric identity yields:

$$I = \frac{1}{2} \cos(4\pi f_s t + \varphi_m(x)) + \frac{1}{2} \cos(\varphi_m(x)) \quad \text{and} \quad (8.79)$$

$$Q = \frac{1}{2} \cos\left(4\pi f_s t + \frac{\pi}{2} + \varphi_m(x)\right) + \frac{1}{2} \cos\left(\frac{\pi}{2} + \varphi_m(x)\right) \quad (8.80)$$

If a low-pass filter is used after both multipliers to provide sufficient attenuation to the signal at a frequency of  $2f_s$ , then the remaining signals are

$$I = \frac{1}{2} \cos \varphi_m(x) \quad \text{and} \quad (8.81)$$

$$Q = \frac{1}{2} \cos \left( \frac{\pi}{2} + \varphi_m(x) \right) \quad (8.82)$$

The phase shift due to the optical path differences is then equal to:

$$\varphi_m(x) = \arctan \left( \frac{Q}{I} \right), \quad (8.83)$$

In order to derive a correct incremental phase shift  $\Delta\varphi_m(x)$ , related to an incremental displacement  $\Delta x$  one has to ensure that the sign of the input values is taken into consideration in order to place the angle in the proper quadrant at the start and the end of the increment. The last step needed is an unwrapping function which properly adds or subtracts  $2\pi$  for each successive  $2\pi$  phase jump that has passed during the incremental displacement. The phase can then be scaled to the displacement value with Equation (8.76) by knowing the refractive index  $n$ , the interferometer constant  $N$  and the centre laser frequency  $f_c$ .

### **Important remark:**

These requirements, regarding the right quadrant and perfectly counting all  $2\pi$  phase jumps, point clearly to the fully incremental nature of practical laser interferometer systems. They all require an initialisation step to set the position counters and phase pointers to zero.

It is not possible to measure absolute positions along the total trajectory, like with the separate marks of encoders. To solve this problem, research is done on interferometry with multiple frequencies, so called *frequency combs* that are based on a laser with short pulses of only a few hundred femtoseconds ( $10^{-15}$ ). This research has not yet resulted in a practical commercial implementation.

### **Velocity limitation of heterodyne interferometry**

An important limitation to the applicability of heterodyne interferometry is due to the maximum value that can reliably be realised for the split frequency of the applied laser beam. When moving in a direction that increases the frequency difference at the measurement detector no real problem will occur as long as the detector can handle the increased frequency level. A movement in the opposite direction, however, will first reduce the frequency

difference to a lower frequency depending on the velocity. When increasing the velocity, at some point the measured frequency difference becomes equal to zero. Above that velocity, the frequency difference will increase again but with a different sign. It is clear that in such a situation all directional information is lost and the measurement system should be reset again.

The minimum split frequency is determined by Equation (8.73):

$$f_s > f_c \left( \frac{N n v_m}{c} \right) \quad (8.84)$$

The interferometer constant  $N$  was equal to two in the previous example but in the following section, examples will be shown with  $N = 4$ . Although a higher interferometer constant will increase the sensitivity and accuracy it requires a corresponding increase of the split frequency.

This property of the heterodyne interferometer forces a choice between accuracy and maximum velocity and has stimulated the development of acousto-optic laser sources with a large frequency difference combined with high-speed signal conditioning and signal processing electronics.

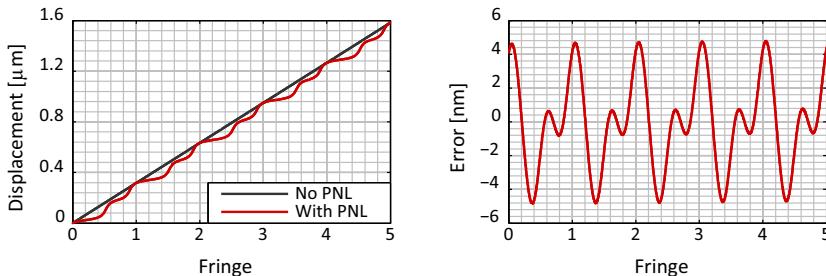
### 8.8.2.3 Measurement uncertainty

There are two major sources of uncertainty in laser interferometer position measurement systems. The first is related to the question whether the interferometer detects only the desired displacement changes. The second uncertainty source is related to the phase measurement and whether it converts only the desired phase shift into a measurement signal.

There is a subtle, yet important difference between these two uncertainty sources. The former pertains to the interferometer design and direct compensation or cancellation by balancing the system whereas the latter deals more with the laser system employed, measurement environment, phase-to-displacement conversion, and sample properties.

The uncertainty factors related with mechanics will be presented in Section 8.8.3 at the end of this chapter. In the following, first the internal errors of the interferometer will get attention:

- **Periodic errors**, causing a repetitive modifying additional measurement error at the nanometre and sub-nanometre levels.
- **Frequency stability**, that acts as a modifying error.
- **Wavefront errors and pointing stability** due to non-ideal optical parts.



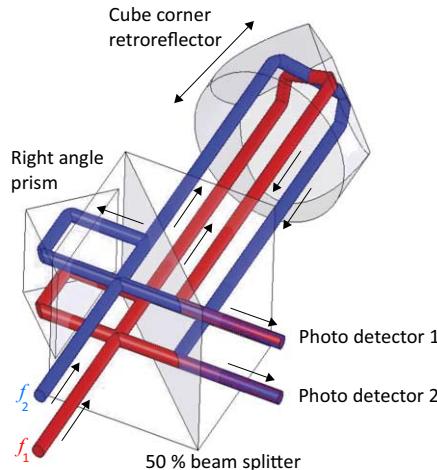
**Figure 8.80:** Comparison between a linear signal and a signal with periodic nonlinearity, highly exaggerated for illustrative purposes. The second graph shows a linear displacement with 5 nm of periodic errors with the nominal slope removed. The periodic errors typically have a first-order component with one cycle per fringe and a second-order component with two cycles per fringe. Higher-orders may appear from additional ghost reflections or stray signals.

(Courtesy of Jonathan Ellis)

- **Refractive index changes** due to the properties of the medium between the interferometer and the moving object.

### Periodic errors

The working principle of a heterodyne interferometer is based on the capability to fully separate the measurement and reference beam by means of their polarisation. Unfortunately the polarisation direction is never fully orthogonal due to imperfections in the source or the polarising optics. This means that there is always a small part of the reference beam that travels along with the measurement beam and the other way around. This unwanted effect is called *polarisation mixing* and causes small changes in the measured phase between the measurement interference signal and the reference signal that is a periodic function of the displacement in relation to the wavelength expressed in fringes as were introduced with the homodyne interferometer. One fringe is equal to a  $2\pi$  phase shift corresponding to a displacement of  $x = \lambda/N$ . These phase differences induce a deterministic *periodic error* as shown in Figure 8.80 that can in principle be compensated by means of software and calibration as long as the conditions don't change. This compensation can only be as good as the stability of the system and in some cases it is better to avoid them by fully separating both beams directly from the beginning.



**Figure 8.81:** Heterodyne interferometer without periodic errors. Two parallel beams from the optical source ( $f_1$  and  $f_2$ ) travel to the beam splitter and the reflected beams travel toward the right angle prism. The transmitted beam travels to the retro-reflector. Then, the reference and measurement beams can be recombined by the beam splitter to create an interference with opposite phase directions, detected by the photo detectors.

(courtesy of Ki-Nam Joo)

Figure 8.81 shows an example of such a system with one degree of freedom. Two separate laser beams with a fixed split frequency, originating from an acousto-optic modulated laser source are inserted in a 50 % beam splitter, similar as the one used in a homodyne interferometer. One half of both beams is reflected as two separate reference beams by a right angle prism, back towards the beam splitter and the other half is transmitted as two separate measurement beams towards the movable cube-corner retro-reflector. The right angle prism has line symmetry which means that the rays are reflected in the same plane orthogonal to the edge line of the prism as where they entered. As a consequence the lower reference beam  $f_1$  will stay at the lower side of the interferometer and the higher reference beam  $f_2$  will stay at the high side. The point symmetry of the cube-corner causes both beams to change position and the measurement beam  $f_1$  arrives at the beam splitter at the same location as the reference beam  $f_2$  while the measurement beam  $f_2$  arrives at the same place as the reference beam  $f_1$ .

These two beams both interfere at the beam splitter and their interference signals will show an opposite phase shift as function of the measured dis-

tance between the interferometer and the moving retro-reflector.

The first advantage of this configuration is that no polarisation mixing can occur before the interference takes place as periodic errors are related to the changing trajectory of the measurement beams to the moving retro-reflector. The second advantage is the double sensitivity with interferometer constant  $N = 4$  because of the fact that four beams are travelling the same trajectory to the moving retro-reflector. The last advantage is that no use has been made from the polarisation direction. This enables its application in more complex interferometers where the polarisation direction can be used for other reasons, like for creating a plane-mirror version of this interferometer along the line of thoughts that are presented in the next section.

### **Frequency stability**

The uncertainty of the interferometer measurement itself is fully based on the phase measurement that is equal to Equation (8.76):

$$\Delta\phi_x = \pm \frac{2\pi f_c N n}{c} \Delta x \quad (8.85)$$

The first source of uncertainty is the frequency of the light  $f_c$ . In principle lasers can be made with a stable frequency up to  $10^{-11}$ , or 10 pm per metre, by thermal stabilisation and other methods, which is sufficient for most measurements. Further all factors are constant with the exception of the refractive index  $n$  that is related to the propagation speed in the environment where the measurement takes place, which is mostly in air.

### **Wavefront errors and pointing stability**

Interference of the reference and measurement beam takes place over the total overlapping area of both beams. As long as both beams have a flat and parallel wavefront the interference is identical over the total surface but any deviation to that ideal flat wavefront will cause a phase shift and reduction of the total observed interference at the photo detector. With the exception of the purposely applied tilt in the wavefront of the homodyne interferometer to create moving fringes, normally all wavefronts in a laser-interferometer measurement system need to be parallel and without any exception always the wavefronts need to be flat. This poses extreme requirements to both the spatial coherence of the laser source, the flatness of the optical surfaces and perfection in the retro-reflectors. In principle *optical flats* are very difficult to manufacture with an exponential cost increase as function of

size and maximum allowable surface topology deviation. The deviations can be expressed using the Zernike modes as described in Section 7.5.2.1 and have to remain at least below  $\lambda/20$  to get an acceptable interference signal but when lateral beam displacements are present, the deviations in the surface need to remain below the required measurement accuracy of the interferometer system.

Next to non-parallelism, a beam that is not pointing to the right direction will induce a so called *cosine error* that is proportional to the cosine of the angle of the beam with the reflecting surface of the measurement mirror. This error is minimal when the direction is orthogonal. When initially all interferometers are mounted such that the measurement beam is as good as possible aligned, any change in the pointing due to temperature or to other mechanical instabilities will have an influence on the measurement error.

### Refractive index changes

The refractive index of air is influenced both by pressure and by temperature and the related measurement uncertainty is a more difficult error to overcome. This is in a large part due to the systematic uncertainty in the equations used to calculate the refractive index. Even with ideal environmental parameter measurements, there still is an uncertainty due to the equations used and the empirical data from which they are based. The typical calculation for refractive index is done with the *modified Edlén* equation, named after the Swedish physicist Bengt Edlén (1906 – 1993). After his original definition it was improved by several scientists in an international comparison of interference air refractometers with participants like Piet Schellekens of the Eindhoven university of technology and Jo Spronck from Van Swinden laboratories who presently is active in metrology at our University laboratories in Delft. Other participants included G. Wilkening and F Reinboth from PTB Germany and K.P. Birch and M.J. Downs of NPL in England. This highly accurate experimental work resulted in a publication in Metrologia 1986,22 and a successor paper by the two scientists of NPL in Metrologia 1993,30 from which the following equation is taken for the refractive index difference of air relative to vacuum:

$$(n - 1)_{T,P} = \frac{P(n - 1)_\sigma}{96095.43} \frac{1 + 10^{-8}(0.601 - 0.00972T)P}{1 + 0.003661T} \quad (8.86)$$

The dispersion factor  $(n - 1)_\sigma$  is equal to:

$$(n - 1)_\sigma = \left( 8343.05 + \frac{2406294}{130 - \sigma^2} + \frac{15999}{38.9 - \sigma^2} \right) \cdot 10^{-8} \quad (8.87)$$

The wave number  $\sigma$  equals the inverse of the wavelength in vacuum ( $1/\lambda$ ). These equations give the refractive index  $n_{T,P}$  for dry air as a function of the wave number  $\sigma$  in  $\mu\text{m}^{-1}$ , the temperature  $T$  in degrees Celsius and the pressure  $P$  in Pascals for a wavelength in the range of 350 – 650 nm.

The refractive index for non-dry air  $n_{T,P,P_v}$  is calculated by:

$$n_{T,P,P_v} - n_{T,P} = -P_v(3.7345 - 0.0401\lambda^{-2}) \cdot 10^{-10} \quad (8.88)$$

where the vapour pressure  $P_v$  is in Pascals.

Using these equations, the traceable uncertainty in the calculated refractive index is not better than approximately one part in  $10^8$  because of the uncertainty in the measurements of pressure, temperature and humidity. This value is equal to a traceable uncertainty of 10 nm over a metre, related to agreed standards with long term stability demands. Fortunately this uncertainty contains unknown but rather constant errors. When only short term relative measurements have to be done in the order of minutes or seconds like is often the case in mechatronic positioning systems, the uncertainty can be as low as 100 pm.

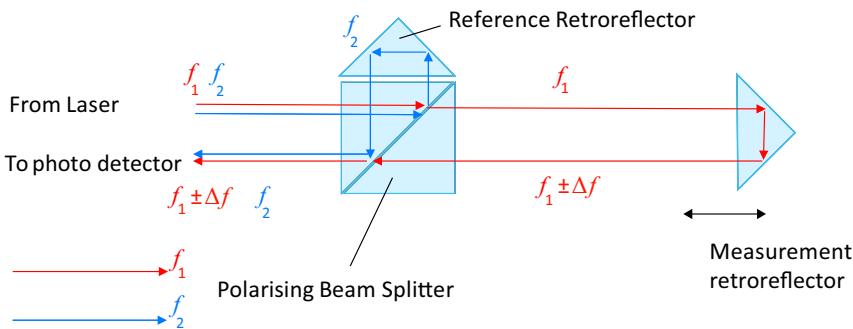
By taking the partial derivative of the modified Edlén equation to temperature, pressure and humidity the individual sensitivity for these modifying error inputs can be estimated in a certain environmental condition setting. With an atmospheric pressure of  $10^5$  Pa, a temperature around  $20^\circ\text{C}$  and a wavelength of 600 nm, the equations show the following approximated partial derivatives in  $[\text{ }^\circ\text{C}^{-1}]$  and  $[\text{Pa}^{-1}]$ :

$$\frac{\partial n_{T,P,P_v}}{\partial T} \approx 1 \cdot 10^{-6}, \quad \frac{\partial n_{T,P,P_v}}{\partial P} \approx 2.8 \cdot 10^{-9}, \quad \text{and} \quad \frac{\partial n_{T,P,P_v}}{\partial P_v} \approx 3.6 \cdot 10^{-10} \quad (8.89)$$

This means that the traceable uncertainty of one part in  $10^8$  approximately corresponds with environmental changes of 3.7 Pa, 10 mK and 27 Pa. A partial vapour pressure of 27 Pa is about 1 % of the saturation vapour pressure of water of  $2.4 \cdot 10^3$  Pa. It is clear that of these factors the temperature has the largest effect and a temperature difference of only 1 mK already gives an error of 1 nm over 1 metre of optical path length.

With slow moving systems in a thermally controlled environment it is often possible to determine the refractive index of air at only one location inside the machine by means of a *wavelength tracker*. This element consists of a separate interferometer with a measurement retro-reflector at a fixed mechanical position relative to the interferometer. Any measured phase changes of this interferometer can then only be caused by refractive index changes of the air by the temperature, pressure and humidity.

Fast moving machines however create so much turbulence inside the machine that next to an accurate thermal control, also the airflow around the



**Figure 8.82:** Basic single-pass heterodyne interferometer with two cube-corner retro-reflectors to shift the beams and prevent sensitivity for tilt.  
Note: The rays are shown shifted and the positions are only indicative. The polariser before the detector is not shown.

moving object needs to be controlled. The air needs to flow faster than the maximum velocity of the moving object in order to avoid mixing of air with different temperatures due to warm objects like actuators.

This airflow necessitates fast blowing *air showers* with temperature stabilisation of better than 1 mK for nanometre range errors. The additional complication of these measures, on top of other requirements like the maximum allowable vibration forces exerted by the turbulent air, is one of the reasons for applying encoders instead of laser interferometers in these applications.

When working in vacuum, these problems do not occur and under that condition laser interferometers are well suited for the most critical requirements, for instance with stages for EUV lithography wafer scanners. Also in space instrumentation distance measurements with laser interferometers is often the best solution, especially when long distances between satellites have to be measured.

#### 8.8.2.4 Different configurations

Several practical configurations are applied in laser interferometry. The examples shown all apply to heterodyne interferometry because of their application in complex multi-axis measurement systems and the possibility to direct the beams by virtue of their different polarisation.

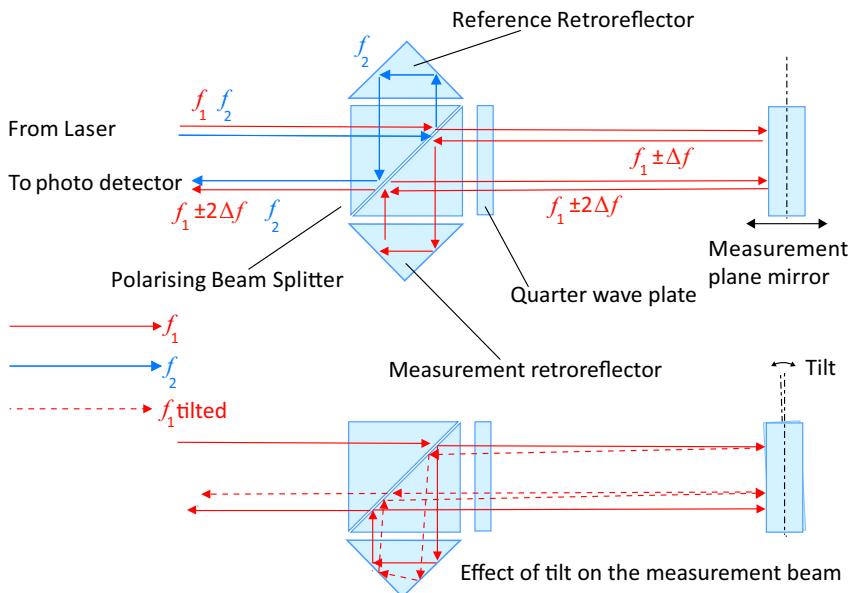
The first example from Figure 8.82 is the most simple heterodyne interferometer configuration possible. It is a single-axis single-pass interferometer and instead of applying  $\lambda/4$  plates to direct the light to a detector at another

location the detector is placed aside of the laser source. To prevent colliding laser beams it uses two cube-corner retro-reflectors.

The light from the source enters at the upper half of the polarising beam splitter where one of the frequencies is reflected and one is transmitted based in their polarisation direction. One cube-corner serves as reference mirror and shifts the beam to the other half of the polarising beam splitter. The other cube-corner serves as measurement mirror. It prevents wavefront errors that would otherwise be caused by angular movements and also shifts the measurement beam to the other half of the polarising beam splitter. In this way the return path does not collide with the entry path and a detector can be placed next to the laser source.

It should be noted that instead of a cube-corner also a right angle prism can be used in this configuration for the reference beam. The main benefit of a cube-corner in this configuration is its insensitivity for rotation in any direction when mounting the element. A right angle prism needs to be mounted sufficiently well that the reference beam will be pointed at the same location as the measurement return beam.

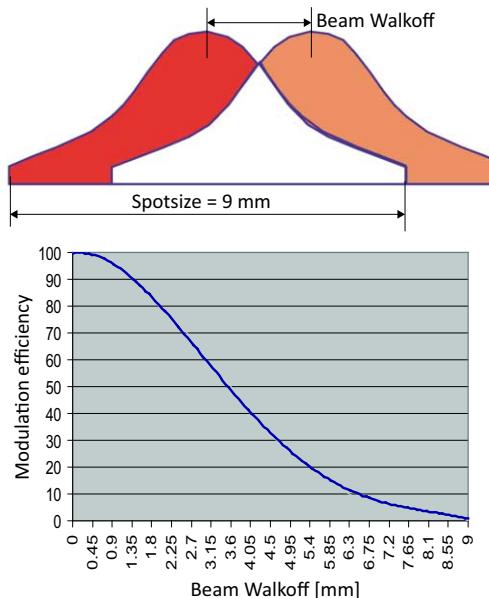
The measurement retro-reflector always needs to be a cube-corner to avoid shifting of the beam and the creation of non-coplanar wavefronts due to angular motions.



**Figure 8.83:** Basic dual-pass heterodyne interferometer with two cube-corner retroreflectors to shift the beams and a plane-mirror as moving element. One  $\lambda/4$  plate changes the polarisation of the measurement beam, causing it to be reflected at the polarising beam splitter and follow a second trajectory to the moving mirror. The lower image shows the effect of tilting of the measurement mirror on the trajectory of the measurement beam. Also here the polariser before the detector is not shown.

### Dual-pass plane-mirror interferometer

The dual-pass plane-mirror interferometer as shown in Figure 8.83 has been developed to enable measurements in more directions with multiple interferometers by replacing the cube-corner with a plane mirror. When the mirror is large, this configuration allows a displacement of the mirror in a perpendicular direction to the measurement direction without affecting the measurement. This configuration uses one  $\lambda/4$  plate and two cube-corners at the interferometer. The reference beam follows the same path as with the single-pass interferometer but the measurement beam follows a different path. After the polarising beam splitter its polarisation is changed from linear to circular by the  $\lambda/4$  plate and after reflection to the moving mirror its polarisation is converted again in a linear polarisation by the  $\lambda/4$  plate but now in the orthogonal direction. As a consequence, the measurement beam



**Figure 8.84:** Tilting of the measurement mirror in a dual-pass plane-mirror interferometer causes the measurement beam to shift. As a consequence the measurement beam and the reference beam will not completely overlap, called “beam walkoff”. Interference takes place only in the overlapping regions.

will be reflected by the polarising beam splitter and the second cube-corner and is directed a second time towards the moving mirror. In this second trajectory, its polarisation will be converted via circular to the original linear polarisation direction enabling it to pass through the polarising beam splitter and interfere with the reference beam.

The dual-passing means that a movement of the mirror gives a double phase shift and double Doppler effect on the measurement beam. The related interferometer constant equals four, which is confusing with the general wording of this interferometer as a dual-pass type. The “dual-pass” refers to the fact that always at least the distance between the object and interferometer has to be passed twice in both directions. Because of the interferometer constant  $N = 4$ , the measurement signal will hence be double sensitive to the movement and as mentioned in the previous section this also requires a higher split frequency to be able to perform a high velocity measurement. Next to this increased higher sensitivity for movements in the measurement direction and insensitivity for large movement in the orthogonal directions, also the sensitivity for tilt of the measurement mirror

is zero for small rotations. This can be seen by following the dashed rays in the lower drawing of Figure 8.83. The ray, returning after the first reflection will be tilted with double the tilt angle of the measuring mirror.

The four reflections on the polarising beam splitter and the cube-corner will return the beam for the second trajectory with the same double tilt angle as after the first reflection and the second reflection will add that same amount of tilt but now in the opposite direction. The result is an unchanged optical path length of the measurement beam.

Although the optical path length is not changed, the measurement beam to the photo detector shows a lateral displacement of the rays, called *beam walkoff* (BWO).

With the tilt angle  $\vartheta_t$  in radians [rad] and the measurement mirror at a distance  $d_m$  from the interferometer the beam walk off is equal to the following expression:

$$\text{BWO} = 4 \cdot d_m \vartheta_t, \quad (8.90)$$

Because of this beam walkoff, tilting is only allowed as long as the rays are all kept inside the interferometer and as long as the reference beam and the measurement beam will sufficiently overlap to induce interference. Figure 8.84 shows the effect of the beam walkoff on the interference efficiency. The part of the beams that do not overlap will give a DC irradiance signal to the sensor and with increased beam walkoff the AC interference part will become relatively smaller.

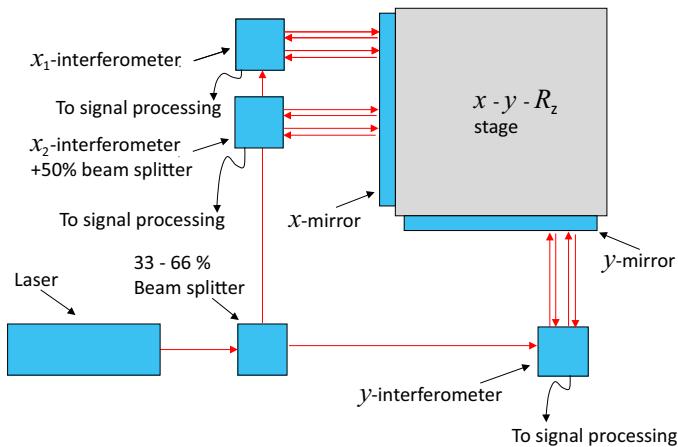
By increasing the diameter of the beams a larger tolerance on tilt is allowed, forcing the application of large optical parts. Because of the optical requirements on flatness and wavefront errors these parts are however extremely expensive.

### Multi-axis laser interferometers

It was mentioned that plane-mirror interferometers allow movements in the orthogonal direction of the measurement direction. This enables their use in measurements of solid objects like a wafer stage of a wafer scanner in more directions.

Figure 8.85 shows the lay out of the three degree of freedom wafer stage from the front page of this book. One interferometer is used for the  $y$  displacements and two interferometers are used for the  $x$  displacement.

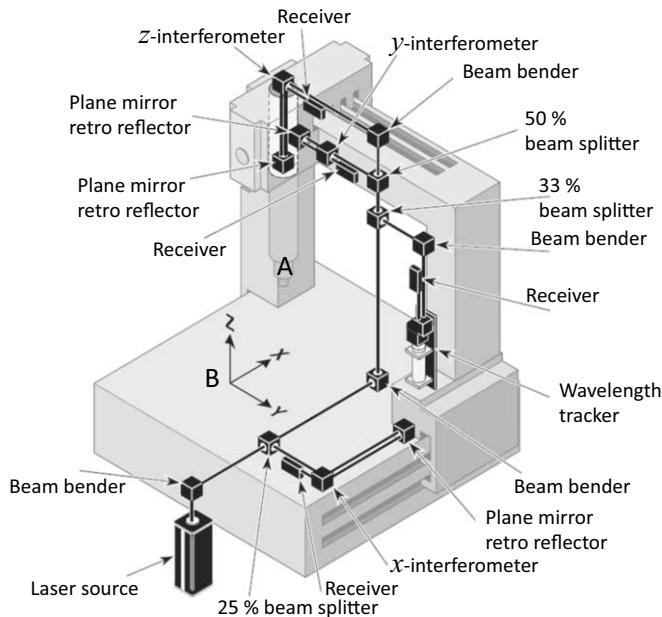
The difference between both  $x$  measurements is a measure for the rotation around the  $R_z$  axis. This rotation can only be small in view of the remarks made previously regarding beam walkoff but still it is sufficient for its use



**Figure 8.85:** With three plane-mirror interferometers measurements can be done in three directions while  $x_1$  and  $x_2$  define both the  $x$  displacement and a rotation  $R_z$  around the  $z$ -axis.

in wafer scanners. By expanding the measurement mirror in the  $z$  direction a full 6-axis measurement system can be realised as will be presented in Chapter 9.

As a last example, the 3-axis  $x - y - z$  measurement system from Figure 8.86 shows the possibilities to direct the light to any measurement location. It also incorporates a fourth measurement branch with a wavelength tracker that enables to determine the refractive index changes of the air by temperature or pressure.



**Figure 8.86:** A three degree of freedom  $x - y - z$  measurement system on a gantry type of machining centre uses different optical components to direct the light to the measurement locations. The wavelength tracker enables to compensate for refractive index variations. The metrology loop from **A** to **B** is very long and indirect.

### 8.8.3 Mechanical aspects

Position measurement is a relative measurement to a known reference. With precision long-range measurements, the reference is often made from thermally solid material mounted on a vibration free environment to prevent any measurement errors that can occur by deformation of the reference. The frame holding this stable reference environment is called a *metrology frame* and this frame represents the “sacred world” of the precision metrologist. In Chapter 9 on wafer scanners, this metrology frame is shown to be solidly connected to the lens in order to guarantee the positioning accuracy of the image to the wafer. Next to this extreme precision application example, also in other applications the metrology frame should be well connected to the reference location, like the position of a work piece in a precision machining centre. With single-axis positioning systems, this metrology frame can be positioned quite close to the measurement location but with multi-axis measurement systems like the gantry type configuration of Figure 8.86 this

is no longer possible.

With this configuration, the metrology frame is not a separate frame but consists of the base and the complete gantry. The relative position measurement from the moving part **A** to the solid table **B** is done in a very indirect way, described by means of the *metrology loop*. The metrology loop is defined by the shortest path that carries information about the relative position of two or more measurement locations and consists of a series of solid objects measurably connected by position measurement sensors or a calibrated sliding mechanism. In the example from the figure, the metrology loop runs from Position **B** via the table to the  $x$ -interferometer that connects its position information to the gantry. The metrology loop runs further through the gantry to the  $z$ -interferometer that connects the position of the gantry to the moving part **A**. All the parts in the loop contribute to the total measurement and as long as all properties of all parts in the entire metrology loop are exactly known, no errors will happen. Unfortunately reality is far from ideal and for instance imperfections in the sliding mechanism and thermal effects on the solid material influence the measurement, often in an unknown and undesired way.

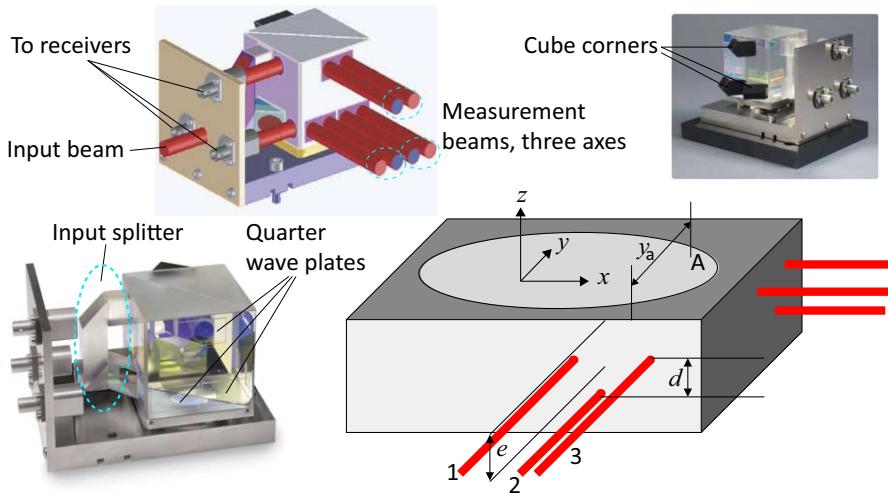
### 8.8.3.1 Abbe error

One important source of errors is based on angular movements combined with long arms. It was the German physicist Ernst Karl Abbe (1840 – 1905) who introduced the term *Abbe error* for these errors. He stated the following:

If errors in parallax are to be avoided, the measuring system must be placed coaxially with the axis along which the displacement is to be measured on the workpiece.

Based on this general statement it was concluded that either the displacement measuring system should be in line with the object of which the displacement is to be measured or the angular motions should be measured separately in order to compensate the effects. An example of such a solution is shown in Figure 8.87. In this case the position of a point **A** on top of the measurement mirror must be measured but it is not possible to direct the laser beams coaxial with this position as the mirror needs to be moved also in the  $z$  direction. In that case additional measurement beams can provide information on the rotation angle around the  $x$ -axis.

It is extremely important to be aware that any error in this angular measurement will be converted into a proportional error in the  $z$  measurement



**Figure 8.87:** By measuring with multiple interferometer beams, the rotation around the  $x$ ,  $y$  and  $z$  axes can be measured. The Abbe error, caused by the non-coaxial measurement with the surface plane  $e$  and a rotation around the  $x$ -axis, can be compensated by the angular measurement based on the distance  $d$  of two of the interferometer beams. A “monolithic” three axis interferometer of Agilent Technologies is shown that can accomplish this measurement in one unit. It integrates three independent plane-mirror interferometers with the high relative stability that is needed for a reliable angular measurement.

direction linearly depending on the distance  $y_a$  to the point of interest. For this reason the requirements on stability are even more extreme than for the linear movements and it is always better, when possible, to measure as good as possible *in Abbe*, as it is called in metrology terms.

The interferometer unit of Agilent Technologies clearly demonstrates the complexity of such an interferometer as all optical parts are fully integrated to guarantee a stable relation between the three further independent plane-mirror interferometers. This integration of multiple optical components can be done by means of adhesives but also direct optical bonding between the polished surfaces is possible when these surfaces are sufficiently flat. Only by such a high level of integration the necessary accuracy in the angular measurement can be achieved.

These principles are applied to the extreme in the wafer scanners of ASML that are presented as the closing application case of this book.

# Chapter 9

## Precision positioning in wafer scanners

In Chapter 1 the invention of the first wafer stepper, the Silicon Repeater of Philips Electronics was memorised. This machine was designed for the purpose of realising the complicated structural lay-out of an integrated circuit and its successors at ASML<sup>1</sup> became an important factor in the worldwide proliferation of electronics, because of its capability to expose an image of a reticle on a wafer with very small details.

### 9.1 Introduction

The smallest details in an integrated circuit are called *features* and their minimum size is mainly determined by the resolution of the optical exposure system as indicated by the Critical Dimension (CD). In Chapter 7 it was explained that the resolution of an optical system is determined by the wavelength  $\lambda$  of the light, the Numerical Aperture (NA) of the lens and a  $k_1$ -factor that mainly is determined by the illumination system. With these variables the critical dimension was shown to be equal to:

$$CD = k_1 \frac{\lambda}{NA} \quad (9.1)$$

This resolution is a theoretical value that is influenced by the surrounding pattern. The wave character of the light creates an airy-disk field profile

---

<sup>1</sup>Several of the figures in this chapter and the background information is graciously provided by **ASML**, market leader in lithographic exposure systems for the semiconductor industry.

of the diffraction limited feature and the “ringing” around the peak will add vectorial to the airy-disk field profile of neighboring features. As a result the CD of isolated features is different from the CD of densely packed features. Several definitions for the CD are used for this reason of which one is chosen to be used further in this chapter, the *half-pitch* CD. Other than with isolated features, the half-pitch CD refers to a regular grating of lines and spaces of which the width of a line equals half the period (pitch). Based on the basic formula for the CD, research and development on these machines has always focused on the reduction of the wavelength  $\lambda$  and the  $k$ -factor and on maximising the NA. The need for a short wavelength has resulted in a continuous development of light sources with ever smaller wavelengths, starting with Mercury arc discharge lamps with 436, 405 and 365 nm wavelength, followed by excimer lasers that produce light with a wavelength of 248 and 193 nm. The next step will be an extreme ultraviolet (EUV) light source with 13 nm wavelength. With EUV light only catoptric optics with mirrors can be used, because no transparent material exists for this short wavelength. The improvements to the  $k$ -factor have resulted in the development of several optical techniques to direct the illumination in such a way that as many as possible diffractive orders of the image are captured in the aperture of the lens as was explained in Section 7.4.4.

It was the required high numerical aperture that originally forced the industry to cover the wafer in parts and not as a whole. Exposing a large surface with a high numerical aperture would only be feasible with extremely expensive lenses with a very large diameter. For that reason it was assumed more affordable to use a smaller lens and expose the wafer in steps. Each separate exposed area is called a *die* and one die can contain a multiple of integrated circuits. In practice the size of a die is determined by the projection lens and with a wafer stepper its size was approximately  $18 \times 18$  mm square. Still for practical reasons the numerical aperture in air is maximised to a value of about 0.93 and because of that limitation, most modern exposure systems use water as intermediate medium with a refractive index of  $\approx 1.44$  at 193 nm. This has increased the practical value of the numerical aperture to  $\approx 1.35$ .

Wafer steppers have remained the main method for the most critical layers with the finest details until circa 1998 when the wafer scanner entered the arena.

The principle of the wafer scanner was introduced much earlier by the American company Perkin Elmer that applied a scanning principle in an exposure system with a magnification of one from reticle to wafer. Although it required a reticle with the same size as the wafer it was a good solution

for the contamination and damage problems of the original exposure principle at that time that used a reticle in direct contact with the wafer. Like with a document scanner, a light stripe exposes the reticle in a scanning motion simultaneous with the wafer. Only in this case both wafer and reticle are moving relative to the stationary light stripe. The wafer scanners of ASML were based on that same scanning principle while they applied the demagnifying lens of the wafer steppers to achieve a better resolution.

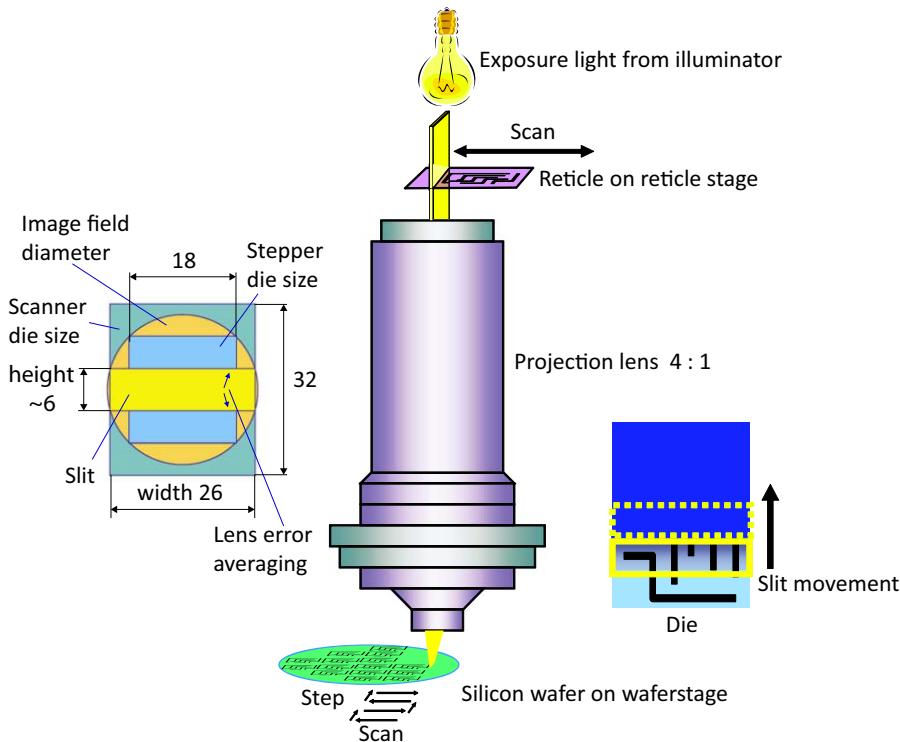
### 9.1.1 The wafer scanner

Figure 9.1 shows in a very schematic way how the exposure of a silicon wafer by a wafer scanner takes place. The illuminator defines a light stripe, called the *slit* with a certain width that is stationary with respect to the projection lens. Instead of using a stationary reticle, in a wafer scanner the reticle is placed on a *reticle stage* that performs a scanning motion, where the illumination system illuminates only one part of the reticle simultaneously. The projection lens images the reticle on the silicon wafer that also is placed on a scanning positioning system, the *wafer stage* that moves synchronously with the reticle stage. The wafer is firmly clamped to a *wafer table* on the wafer stage by means of vacuum. The wafer table is also often called the *wafer chuck* or just “the chuck” because it clamps the wafer tightly to the moving stage. Smaller details can be defined by a relatively large reticle when compared with the scanners of Perkin Elmer, because of the four times demagnification of the projection lens and the exposure of only one die simultaneously. The disadvantage of this demagnification is that the reticle has to move four times as fast with corresponding high acceleration levels and reaction forces. In spite of these dynamic drawbacks this demagnification is necessary because of the complexity to create reticles with sufficient quality. The four times larger details enable the electron beam pattern generators to realise the pattern including the *assist features*<sup>2</sup>, small details that are added to the IC pattern to tweak the imaging and enhance the resolution around corners and in area's where a neighbouring detail is very close by.

The scanning exposure of the reticle has several advantages over the exposure of a stationary reticle as is done in a wafer stepper.

The first advantage of scanning versus stepping is the averaging of small image position errors. Every point on the reticle is exposed over the full

<sup>2</sup>It is beyond the scope of this book to enter deeper in the physics behind these assist features that work on the wave character of the exposed light. The theory on these resolution enhancement methods is a real field of experts working with highly advanced modelling software.



**Figure 9.1:** Basic principle of a wafer scanner exposure system. A small “slit” of light exposes a pattern on a transparent reticle and the pattern is imaged on the wafer by a four times demagnifying projection lens. A simultaneous movement of reticle and wafer results in a full image of the pattern on the wafer. To cover the entire wafer, the scanning motion is repeated with an intermediate stepping motion to the next die position. The size of the die is increased by utilising the full width of the circular image field of the lens and by maximising the scanning stroke.

width of the exposure slit by the scanning motion and small image position errors by the projection lens, present in one part of the slit, will compensate an opposite error in on another part of the slit. If the errors are random they average out.

The second advantage is related to the maximum size of the image field. A lens made of round lens elements will have a circular image field where all parts can be exposed with the same numerical aperture. Exposing a wafer in circular regions is not very practical because in that case parts of the wafer will not be exposed. Although hexagonal dies could fill the entire

surface this method is highly unpractical when such a die has to be filled with rectangular integrated circuits. For these reasons the die always has a rectangular shape. With a wafer stepper the rectangle has to fit in the circular field and a square die will be the maximum surface that fits.

A wafer scanner that uses a slit with a limited height can utilise a larger part of the width of the image field to approximately 26 mm, without increasing the size and cost of the lens. It even allows optimising the optical image quality in the slit area. Further, in theory, the size of the image in the scanning direction could be made infinitely long but the required large reticle would become extremely expensive. For that reason the die size in the scanning direction is limited to approximately 32 mm.

### 9.1.2 Requirements on precision

The critical dimension is one of the main drivers of precision in a wafer scanner. The continuous reduction of the critical dimension is almost like a law of nature determined by *Moore's law*.

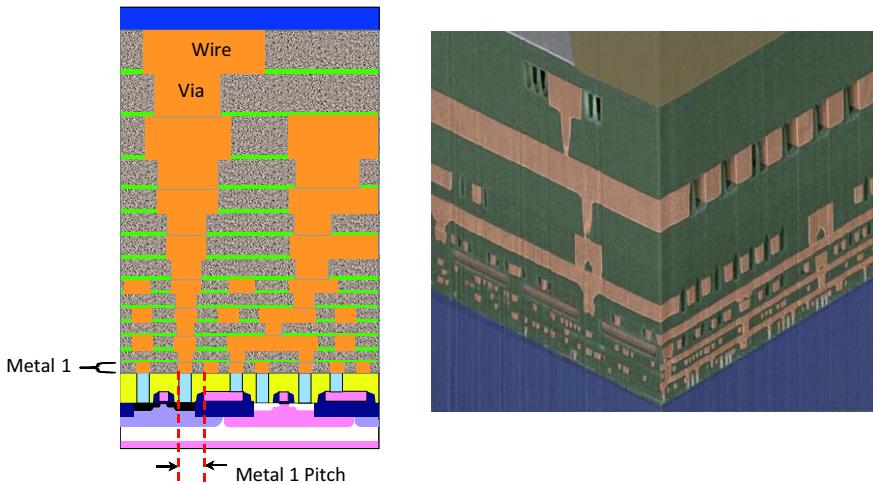
The American chemist and physicist Gordon Earle Moore (1929), co-founder of Intel, defined his law in 1965 based on the observation that the number of transistors on an Intel processor doubled every two years. Although not a real law of nature, this postulation has predicted the continuous exponential growth rate over almost four decades, with a corresponding "shrink" of the smallest details. The industry has organised itself so well in this respect that they even use a jointly defined roadmap, the International Technology Roadmap for Semiconductors ITRS from International Sematech that describes in full detail the expected developments of the different parameters that rule this market for a 15 years forecast.

Next to the critical dimension, the *overlay* is the second main driver for precision. Overlay is the relative position of any pattern layer in respect to the other layers and it should be as small as possible. As can be observed in Figure 9.2, all layers need to be connected by a multitude of small conductive pillars, the *vias*.

Errors in the overlay will impact both the electrical properties of the contacts and the insulation and might even create short circuits.

The maximum allowed overlay value is strongly related to the critical dimension. Originally a maximum overlay in the order of 30 % of the critical dimension was sufficient but recently the overlay requirements were aggravated to a value below approximately 15 % of the CD.

This more severe relative overlay requirement is related to recently intro-



**Figure 9.2:** Cross section of a modern integrated circuit illustrating the multitude of interconnecting layers that have to be positioned with strict requirements on overlay.

(courtesy of Semiconductor Industry Association. The International Technology Roadmap for Semiconductors, 2009 Edition. International SEMATECH:Austin, TX, 2009.)

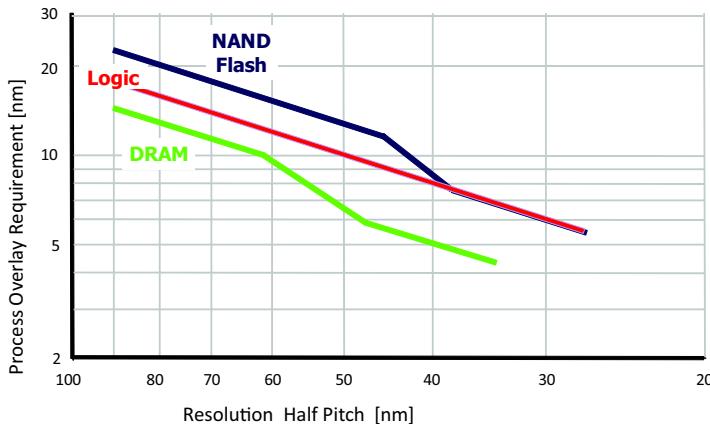
The right image is a CMOS logic microprocessor IC of IBM.

(courtesy of IBM Research)

duced special exposure methods like double-patterning or double-exposure, where lines are exposed in between the lines from a previous exposure cycle. By slightly underexposing a pattern, that can just be imaged by the lens, the developed pattern shows smaller lines than spacings and it becomes possible to expose another set of lines in between. These methods enable IC manufacturers to extend the use of 193 nm wavelength light to ever smaller dimensions, but it has simultaneously resulted in a strong increase in the overlay requirements beyond Moore's law, as can be seen in Figure 9.3.

The third factor of importance for precision is the productivity. The high cost of these machines, mainly driven by the optics, requires an ever increasing speed of operation in order to retain an acceptable return on investment for the IC manufacturer.

The productivity is defined in different ways. The *throughput* of the machine is the easiest measurable item and tells something about how many wafers can be exposed in one hour. On the other hand the productivity tells nothing about how successful the exposures are and so for that reason the number of good exposures per unit of time would be more meaningful for the customer.



**Figure 9.3:** The requirements on overlay follow a trend that outpaces Moore's law. Soon all layers in an integrated circuit need to be positioned well within a few nanometres.

Still for practical reasons the throughput is most frequently used during the design of the machine.

In 2010 a throughput of more than 180 wafers per hour was the state of the art. This value corresponds to 20 seconds per wafer, including loading and unloading. The related extreme velocities and accelerations of the stages and other robotic motion systems cause strong reaction forces with vibration levels that easily could impair the critical dimensions by *fading*, lack of contrast by vibrations during exposure.

All these factors result in a large set of requirements for the wafer scanner. The list in Table 9.1 only gives an overview of a selection of the most important requirements that directly determine the precision positioning-systems, the wafer stage and the reticle stage. The values are approximated as they are only meant to give an idea of the order of magnitude. Note that the wafer stage position error is divided in a low-frequency part, the moving average (MA), causing overlay errors and the high-frequency part, the moving standard deviation (MSD) that causes image fading.

Based on these requirements, this chapter focuses on the following important aspects that determine these extreme performance levels and will pose new challenges for future developments.

- Dynamic architecture, preventing disturbing vibrations of the dynamical sensitive parts.
- Zero stiffness stage actuation with Lorentz actuators to fulfill the demands on dynamic performance.
- Indirect relative position measurement between image and wafer with alignment marks and long-range sensors.
- Motion control, with both feedforward and PID-feedback control with a high level of predictability of the plant dynamics.

**Table 9.1:** Main requirements on precision of the positioning systems in a wafer scanner.

Requirement	Approximate value
Critical dimension (CD).	< 40 nm
Overlay.	< 4 nm
Wafer stage velocity, stepping.	> 2 m/s
Wafer stage velocity, scanning.	> 0.5 m/s
Reticle stage velocity.	> 2 m/s
Wafer stage acceleration.	> 30 m/s <sup>2</sup>
Reticle stage acceleration.	> 120 m/s <sup>2</sup>
Wafer stage metrology error.	< 0.5 nm over 20 s
Wafer stage MA in-plane position error (overlay).	< 1 nm
Wafer stage MSD in-plane position error (fading).	< 10 nm
Focus error.	< 100 nm
Settle time.	< 10 ms

## 9.2 Dynamic architecture

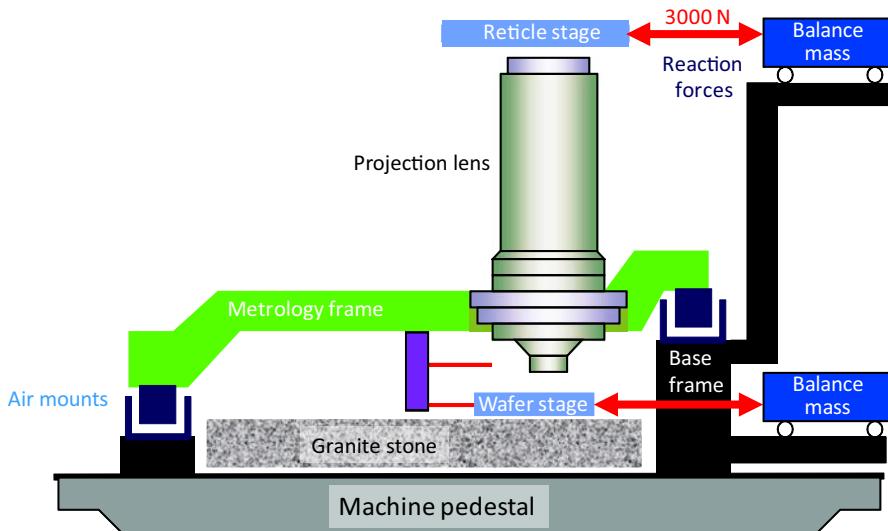
One major rule in precision engineering is to first prevent problems by fully mastering the open-loop dynamics of the mechatronic system before attempting to control its behaviour. For this reason, the dynamic architecture of a wafer scanner, as shown in Figure 9.4, is focused on keeping all non deterministic dynamic disturbances as good as possible separated from the optical imaging system.

The first sources of disturbances originates within the wafer scanner itself due to the large accelerations of the stages. Another important source of disturbing forces are vibrations that are caused by other equipment like the large air conditioning, purifying and processing equipment in the *wafer fab*, the usual name for a semiconductor factory.

These vibrations are reduced by a well designed dynamic structure consisting of the mechanical frames, the use of a *balance mass* to absorb reaction forces and a *vibration isolation system* to protect the most sensitive parts.

The basic mechanical structure of a wafer scanner consists of several parts. The first part is the *pedestal*, a heavy and stiff structure that connects the wafer scanner with the floor of the wafer fab. This floor is typically made from large steel bars and is not extremely stiff relative to the mass of the wafer scanner. The pedestal is often made from solid concrete and is located as low as possible on the fab floor. In this way the pedestal effectively grounds the wafer scanner on the compliant floor structure and reduces the impact of vibrations from inside the machine to the wafer fab.

All the modules of the wafer scanner itself are built on a rigid *base frame* that is made of steel. This main structure of the wafer scanner is directly mounted on the pedestal without additional vibration isolation measures. The projection lens is firmly held in the right position by the *metrology frame* that defines the measurement reference of the stages to the image of the die. The metrology frame is connected to the base frame by means of three *air mounts*, air cushion springs with a very low compliance that serve to reduce the transmission of base-frame vibrations to the metrology frame. The stages are moving in six degrees of freedom and are supported by the base frame, either on air-bearings or by means of an active magnetic support. The support of the wafer stage in the vertical direction often consists of a large air bearing, the *air foot*, that floats on the flattened surface of a large granite stone. Granite is a very stable material and often used for this purpose in precision machines.



**Figure 9.4:** The dynamic architecture of a wafer scanner is based on two principles. The first is to create a vibration free environment for the projection lens and metrology frame and the second is to reduce the vibrations to this “sacred reference” by balance masses and a well-tuned vibration isolation system. In reality the balance mass of the wafer stage includes the granite stone that supports the wafer stage.

### 9.2.1 Balance masses

The horizontal directions need to move with very high accelerations while the vertical direction with the rotations have less stringent requirements on acceleration. These high horizontal acceleration levels of the heavy stages create reaction forces in the order of several kilo Newton. In the first generations of wafer scanners, these reaction forces were directed towards a heavy *force frame* that was mounted on the base frame separately from the other more sensitive parts of the machine.

Although in that way it was expected that the mass of the force frame would absorb these forces, the coupling of the resulting movements to the base frame had caused many problems in practice. The remaining vibrations sometimes excited the dynamics of the fab floor and could disturb any other equipment that was located nearby and they still reached the sensitive parts inside the machine by transmission via the base frame through the air mounts.

For that reason, it was decided in more recent generations of wafer scanners

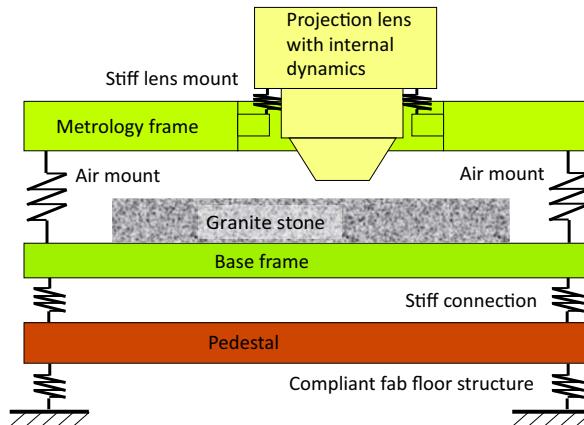
that the reaction forces of the high acceleration  $x$  and  $y$  movements should be absorbed by very heavy balance masses that are not rigidly connected to the base frame but horizontally guided on air bearings or compliant mechanisms.

The balance masses consist of a seismic mass that is directly connected to the stator of the linear actuator that drives the stage. The actuation force from the actuator drives both the stage and the balance mass in opposite directions and causes the relative movement between the mover and the stator of the actuator to become larger than would be the case when the stator was connected to the force frame.

The first consequence of this increased relative movement is an increase of the induced EMF over the motor coils, requiring a higher maximum voltage of the power amplifiers. The second consequence is the need for a longer movement range of the linear actuator with the associated cost of larger coils or magnet assembly.

By increasing the mass of the balance mass its maximum displacement decreases proportionally and for that reason the mass is chosen much larger than the mass of the stage. With a stage of 80 kg the balance mass can have a mass of 500 kg or more. A large mass is also large in volume and measures should be taken to guarantee that the balance mass will not touch any other part of the machine during its movements. To guarantee the consistent free movement of the balance mass measures are taken to keep its average position in the middle of the moving range, compensating the effects of gravity and other non alternating forces working on the balance mass. These measures either consist of a separate actuator with a low stiffness  $k_p$  feedback control or of a passive spring with a sufficiently low compliance to prevent excess transmission of vibrations to the base frame. The method to absorbing reaction forces in a balance mass requires the forces to act as good as possible in a horizontal plane through the centre of mass. Any deviation from this ideal situation will result in a proportional torque around the horizontal axes through the centre of mass with corresponding vibration forces in the vertical direction to the base frame.

A well designed structure with balance masses has helped keeping the transmission of internal vibrations to such a moderate level that the vibration isolation system is able to further reduce these vibrations to the level that is required to achieve the maximum imaging quality of the wafer scanner.

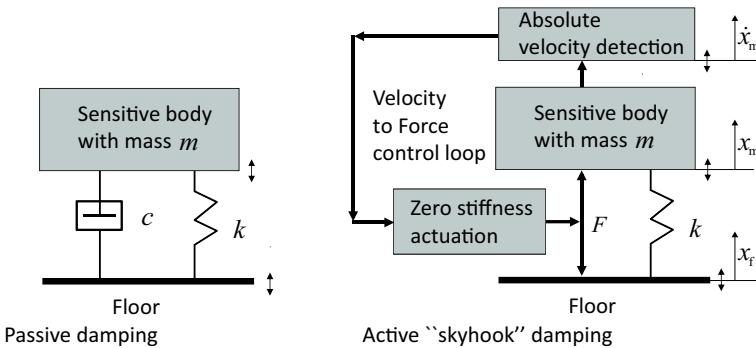


**Figure 9.5:** The frames and springs that determine the sensitivity for external vibrations. The transmission of vibrations from the fab floor takes place via three mass-spring systems in series. The stiffness of the air mounts and the mass of the metrology frame and lens determine the low-frequency limitation of the main vibration isolation element. The stiffness of the connection between the lens and the metrology frame determines a resonating eigenmode that will increase the sensitivity for vibrations at the corresponding eigenfrequency.

## 9.2.2 Vibration isolation

Figure 9.5 shows a rigid body diagram of the parts of the wafer scanner that determine the sensitivity for external and internal vibrations.

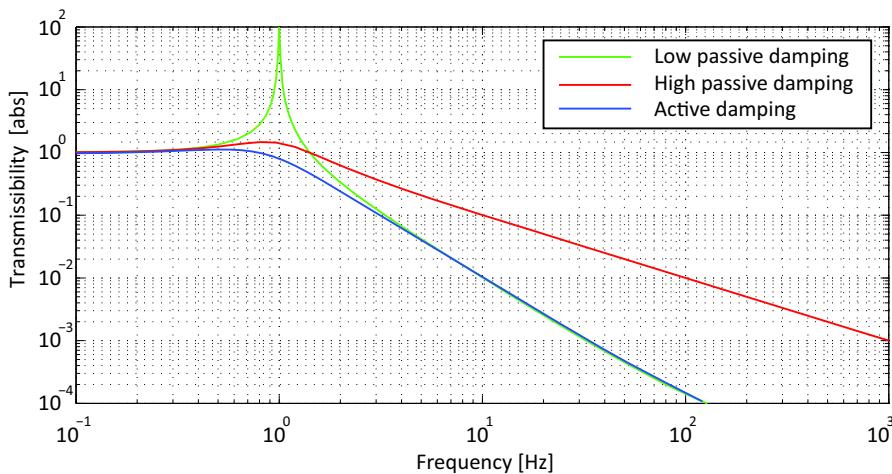
The main line of defence against disturbing vibrations is determined by the transmissibility transfer function of the three air mounts with the combined mass of the metrology frame and the lens. In principle the air mounts consist of air cylinders with a large volume of which the pistons are guided by means of air bearings. These bearings have a sufficiently small air gap that it is not necessary to apply the usual rubber bellows that would otherwise introduce an increased transmissibility at higher frequencies by the intrinsic damping of rubber. The stiffness value of the three air mounts together can be in the order of  $10^5$  N/m. With a combined mass of the metrology frame and the projection lens of approximately 2500 kg, the natural frequency with these air mounts is in the order of 1 Hz. The relatively low stiffness causes a practical problem in relation to changes in the mass of the isolated body. The total static sag of the spring by the mass of 2500 kg would be equal to a quarter of a metre and a small variation of this mass could lead to an unacceptable change in the position of the projection lens. For that reason



**Figure 9.6:** Active “skyhook” damping in a vibration isolation system avoids the transmission of external vibrations through the damper. An absolute velocity detector measures the velocity of the sensitive body relative to an inert seismic mass and its output is used to exert a force opposite to the velocity.

this position in the vertical direction is actively controlled by adapting the amount of air in the air cylinders to the situation. An increased mass would result in a reduced volume with an increased pressure in the air mount. The volume can be restored by adding sufficient air from an external source by means of an air valve that opens and closes as function of the vertical position of the metrology frame.

The next problem with this system is the need for damping. At the natural frequency of 1 Hz the system will resonate, when excited with that frequency by a force or a movement of the floor. As explained in Chapter 3 on transmissibility, a normal viscous damper like a rubber bellows introduces an unwanted connection, increasing the transmissibility of these vibrations at frequencies above the natural frequency. For that reason a *skyhook* active damper is applied. Figure 9.6 shows the principle that is based on the measurement of the “absolute” velocity, relative to a real quiet reference as if the system is “hooked to the sky”. In practice this quiet reference consists of an elastically suspended seismic mass inside the velocity sensor. A suitable sensor for this principle is the geophone as was described in the previous chapter on measurement, that measures the velocity relative to the seismic mass, but the velocity signal can also be obtained by integrating the signal of an accelerometer. The velocity signal is used to create a proportional damping force, opposite to the velocity, by means of a Lorentz actuator of which the force is only determined by the current and not by the position. As a consequence this actuator has no stiffness that would otherwise add



**Figure 9.7:** The transfer function of the active skyhook damping shows an increased attenuation of the vibrations in a frequency band starting at the eigenfrequency of the mass-spring system.

to the stiffness of the supporting mechanical spring  $k$  and increase the transmissibility.

When writing down the equations of motion the transfer function of the transmissibility can be determined:

$$m \frac{d^2x_m}{dt^2} = -c_s \frac{d}{dt}(x_m) + k(x_f - x_m) \quad (9.2)$$

where  $c_s$  equals the control gain of the velocity loop.

With the Laplace transform of the differentiation, the transfer function of  $x_f$  to  $x_m$  becomes:

$$\frac{x_m}{x_f} = \frac{k}{ms^2 + cs + k} = \frac{1}{\frac{m}{k}s^2 + \frac{cs}{k} + 1} \quad (9.3)$$

With the known terms for the eigenfrequency  $\omega_0$  and damping ratio  $\zeta$  the equation becomes:

$$\frac{x_m}{x_f} = \frac{1}{\frac{s^2}{\omega_0^2} + 2\zeta \frac{s}{\omega_0} + 1} \quad (9.4)$$

This transfer function does not show the differentiating term in the numerator that represented the increased transmissibility in the transfer function of a passive damper and as a consequence the Bode-plot of Figure 9.7 shows

clearly an increase in attenuation of the external vibrations, approaching the level as would be obtained without damping.

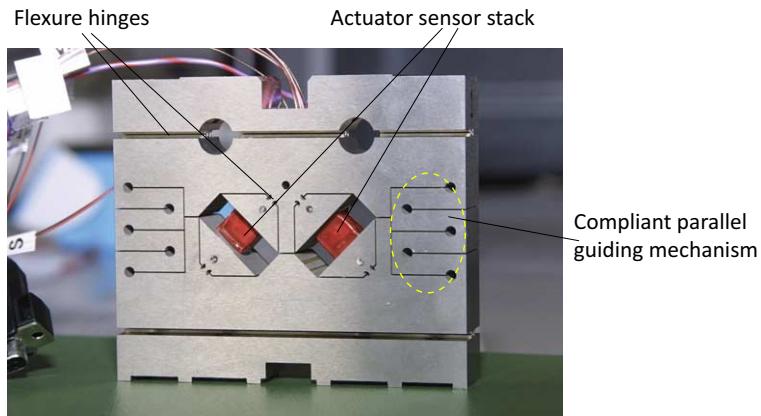
Although this method of active damping is often used, its performance is limited by the noise of the sensor. In principle the sensor needs to detect only very small movements and any noise source will insert an disturbing signal in the damping loop. In practice this reduces the positive effect of the active damping at low frequencies. In reality the power spectral density of the noise is not equal for all frequencies. A piezoelectric accelerometer with a charge amplifier requires an additional integrator to create an inertial based velocity signal. An integrator will amplify the noise at very low frequencies more than at high frequencies due to the feedback capacitor of the amplifier. As was explained in Chapter 8 this is especially problematic with excess noise ( $1/f$ ). Even though a direct inertial velocity sensor like the geophone does not need an integrator, additional amplification at low frequencies might be necessary to achieve a flat velocity response, sufficiently below the natural frequency of the vibration isolation system.

At several research institutes investigations show promising results in the optimisation of the the active part of vibration isolation systems. These investigations include the active reduction of the stiffness of the connection and a virtual connection of the position to a quiet reference. In spite of these developments it is in practice still preferred to base the main part of the vibration isolation based on a heavy mass and a compliant spring because of the low-frequency noise and other dynamic problems in the sensors like cross-coupling.

### 9.2.2.1 Eigendynamics of the sensitive parts

A large mechanical structure, like the metrology frame with the lens, inevitably shows several dynamic eigenmodes with their related eigenfrequencies. Figure 9.5 shows springs between the lens and the metrology frame that represent the stiffness of the mutual connection. Even though this connection is very stiff, it still results in a significant resonance with the heavy lens. Also the optical elements inside the lens with their compliant mechanical mounting determine eigenmodes that might be excited by external vibrations. The related eigenfrequencies are often in a rather low-frequency range around 50 – 100 Hz because of the large masses and for that reason the vibration isolation system should work especially well in that range.

The low amount of damping that generally is present in mechanical mounts normally creates resonances with high  $Q$  levels that increasing the vibra-



**Figure 9.8:** An piezoelectric active lens mount with “smart-disks” that consist of a combined piezoelectric actuator and sensor. By measuring the deformation a damping force can be generated in an active feedback loop. (Courtesy of Jan Holterman UT)

tions with a factor thirty or more. Even with a well controlled vibration isolation system it is necessary to create additional damping to reduce the amplitude at these eigenfrequencies. Damping in a mechanical mounting structure is difficult to achieve with passive means because of the high stiffness of the connection. With the related small movements it is almost impossible to dissipate much energy in viscosity.

### Smart disk

Damping in stiff connections can be created actively with actuators and a suitable control scheme. The *Piezoelectric Active Lens Mount* (PALM) of Figure 9.8 is an example that is used to dampen the eigenmode of the connection of the projection lens with the metrology frame. Three of these systems are required to create damping in six degrees of freedom.

The basic design of this system is the outcome of a research project at the University of Twente by Jan Holterman under the guidance of Rien Koster who invented the *smart-disk* that fulfils the key role in this system.

A smart-disk consists of a combination of a piezoelectric actuator and sensor stacked on top of each other. In principle the sensor measures the forces that act on the stack and the actuator creates a displacement as function of the applied voltage.

A feedback loop can be created from the sensor to the actuator. Integration

of the measurement signal in the loop (I-control) creates an effect as if a damper is inserted in series with the smart-disk according to the following reasoning.

For damping the force needs to relate to the velocity of the vibrating lens. When  $F_d(s)$  equals the force signal from the sensor in the Laplace domain, I-control will generate a displacement  $x_d$  by the actuator:

$$x_d = \frac{k_i}{s} F_d(s) \implies \frac{s x_d}{F_d(s)} = k_i = \frac{1}{c} \quad (9.5)$$

The force by the movement of the lens will generate a proportional velocity of the actuator which is the behaviour of a damper with damping coefficient  $c = 1/k_i$  in series with the controlled smart-disk.

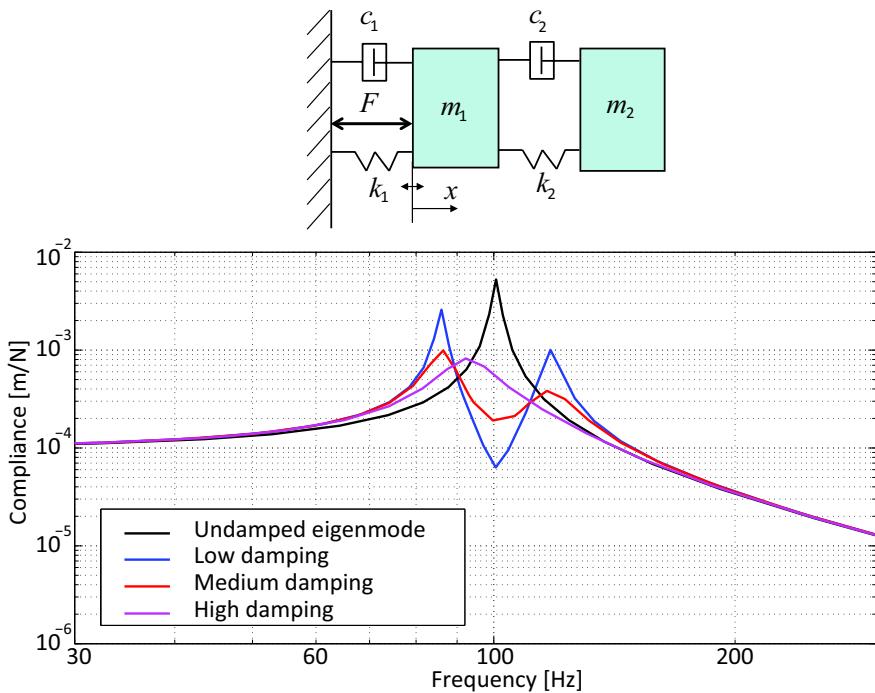
### Anti-resonator or tuned-mass damper

A second method to reduce vibrations due to eigenmodes is by connecting an *anti-resonator*, also called a *tuned-mass damper* to the sensitive body. An anti-resonator is a tuned mass-spring system that is attached to the sensitive body. The decoupling phenomenon of this additional mass-spring system causes a typical combination of an anti-resonance and a resonance in the transfer function from the excitation force to the sensitive body as described in Section 3.3.1 of Chapter 3. The frequency where the transfer function equals zero was shown to be:

$$f_a = \frac{1}{2\pi} \sqrt{\frac{k_a}{m_a}} \quad [\text{Hz}] \quad (9.6)$$

where  $k_a$  and  $m_a$  are equal to the stiffness and mass of the anti-resonator. The anti-resonance frequency  $f_a$  is chosen equal to the undamped eigenfrequency of the lens with its mounting stiffness. The combined transfer function of the anti-resonator and the regular transmissibility for vibration forces from the floor to the movement of the lens via the air mounts is shown in Figure 9.9. The Bode-plot has been derived using the standard equations of motion from Chapter 3 and it shows that the original resonance at 100 Hz of this example is replaced by two resonances with a smaller magnitude depending on the damping of the tuned mass.

The observed shifting of the resonance frequencies is especially useful when the undamped system resonates in a frequency area where external disturbance peaks occur like for instance the 100 Hz hum from high power mains supply units.



**Figure 9.9:** An anti-resonating tuned-mass damper replaces the original undamped eigenmode of the lens with its support by two resonances and an anti-resonance. At a certain value of  $c_2$  the magnitude of the resonances is reduced in comparison with the situation without the anti-resonator.

In those cases where the resonance frequency is not fixed it is also possible to create an actively controlled anti-resonator. By changing the stiffness with proportional feedback, the tuned-mass damper system can be adapted to the right resonance frequency.

## 9.3 Zero stiffness stage actuation

With the explanation of active vibration isolation a Lorentz actuator was used to exert the damping forces. This actuator was chosen because a Lorentz principle shows an almost ideal *zero-stiffness* behaviour, without increasing the transmissibility for vibrations from the vibrating base frame to the sensitive metrology frame.

In a similar way, also the requirements on the accuracy of the stages, with their high level of velocity and acceleration, require the use of these zero-stiffness actuators in order to completely avoid any elastic connection between the moving stage and the surrounding machine parts.

In a fundamental way, the positioning principle of the wafer- and reticle stage is purely based on the second law of Newton, stating that the acceleration of a body is proportional to the force and inversely proportional to the mass of the body. As long as the forces are known and controlled, a body with a known mass can be accurately positioned by a feedback controller as was explained in Chapter 4 with the PD-control positioning of the optical disc readout unit.

In the design of a stage of a wafer scanner most of the effort is invested in full deterministic control of the forces acting on the moving body, consisting either of the wafer or of the reticle on their supporting table. As long as these bodies behave like a rigid body with predictable and reproducible dynamics, the movement of the body is only determined by the forces acting on it. With this reasoning, the term “zero-stiffness” is only related to the mechanical connection to sources of disturbing vibrations and it does not refer to the control stiffness. When zero-stiffness of the actuation is achieved, the body will behave like an ideal inertial mass and the task of the control system is to connect the stages as stiff as possible to the optical system. More precisely stated, the control system needs to connect the reticle stage as stiff as possible to a planned scanning trajectory, relative to the optical system. Simultaneously the control system needs to connect the wafer stage as stiff as possible to the scanning image of the moving reticle.

The wafer stage is mostly taken as example in the following sections but the design principles are equally true for the reticle stage. The main difference is the higher velocity and acceleration with reduced accuracy of the reticle stage in respect to the wafer stage.

### 9.3.1 The wafer stage actuation concept

From all possible actuator types, only the mentioned Lorentz actuator almost ideally complies with the zero-stiffness criterion. Piezoelectric actuators are stiff by definition and reluctance based actuators have a negative stiffness that induces a comparable, however inverted, transmission of vibrations as a positive stiffness.

Unfortunately a simple Lorentz actuator offers only a small linear range. In order to solve this limitation, electronic commutation was introduced in Section 5.2.5 of Chapter 5. The shown example of Figure 5.23 had a long set of three moving coils with a large magnet assembly and in spite of the electronic commutation the range was still rather limited because the coil set and magnet assembly both needed to have the same length in the motion direction.

#### 9.3.1.1 Wafer stepper long-range Lorentz actuator

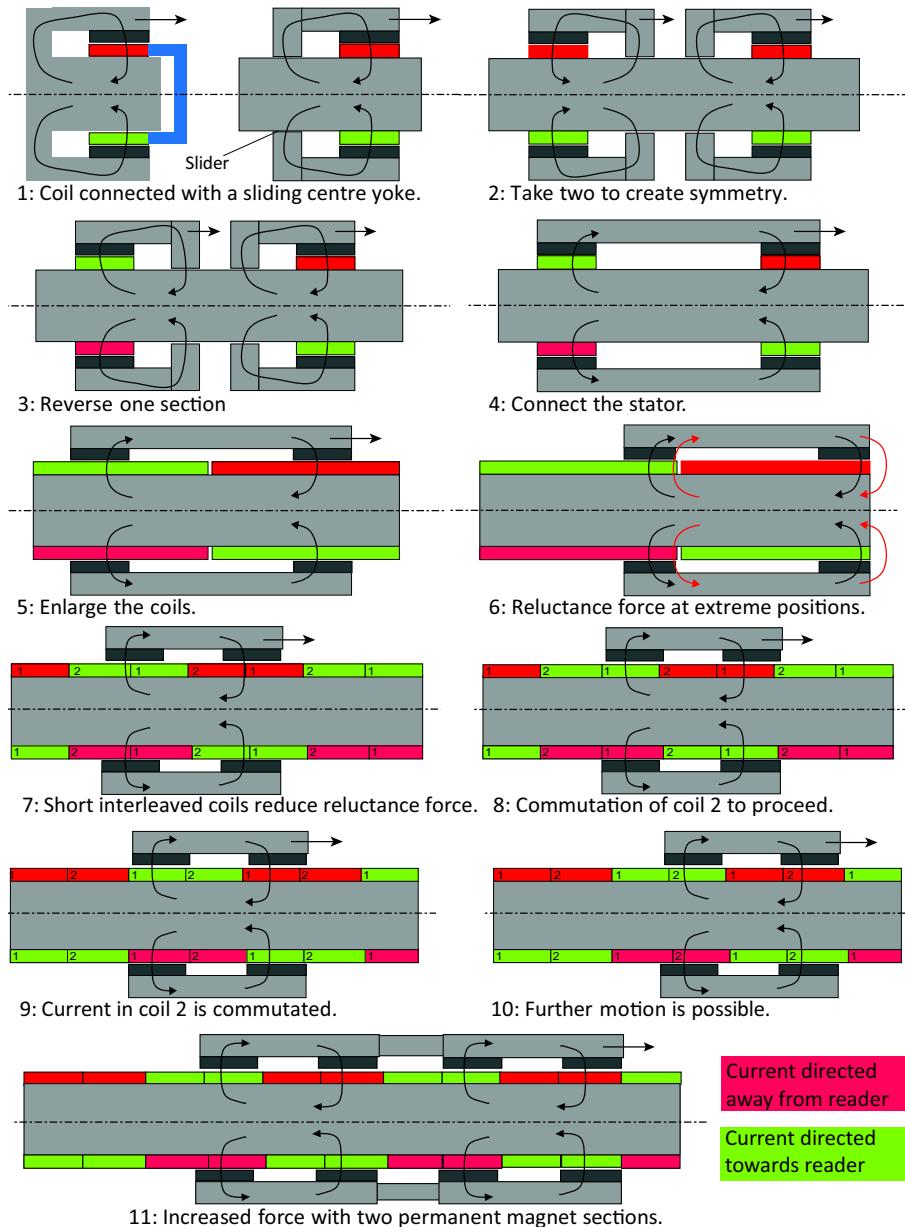
With the design of the first wafer stepper an alternative two phase electronic commutated Lorentz actuator was designed by one of the authors of this book at Philips Research laboratories. This extended-range Lorentz actuator was able to work without the above mentioned limitation and was applied in the wafer stages of all wafer steppers until the introduction of the wafer scanner that required a better solution.

Figure 9.10 shows the thinking steps that have led to this design. It starts with a standard loudspeaker type Lorentz actuator with the permanent magnet system as moving part.

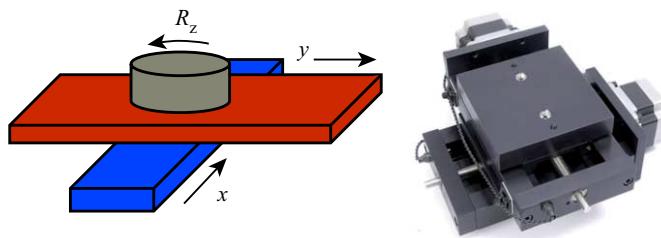
The first transformation step (1) is to divide the moving ferromagnetic yoke in a moving outer part that slides over the cylindrical inner part that holds the coil. This step reduces the moving mass and is allowed as still the permanent magnet field is moving in respect to the coil, creating a  $d\Phi/dx$ . The sliding could be an air bearing but in the next steps the slider will be avoided.

Step (2) is the addition of an equal permanent magnet-coil set around the common central yoke. This step increases the force at a certain current level and becomes even more useful when the direction of both the permanent magnetic flux and the current in the coil of one of the actuators is reversed in step (3).

The sliders are deleted in the step (4) because after step three the flux of both permanent magnets can be shared. This step is essential as now there is a symmetrical set of two permanent magnets connected by one simple



**Figure 9.10:** Design of a zero-stiffness Lorentz type actuator with long linear-motion range by electronic commutation. The direction of the current in the coils corresponds with a force on the moving part to the right. The curved arrows indicate the direction of the permanent magnetic flux.



**Figure 9.11:** The usual way of creating a three-axis positioning system is done by stacking three single-axis stages on top of each other. The driving forces are not directed to the centre of mass, resulting in unwanted torques and rotations. For that reason this configuration is not suitable for fast stages with a high precision.

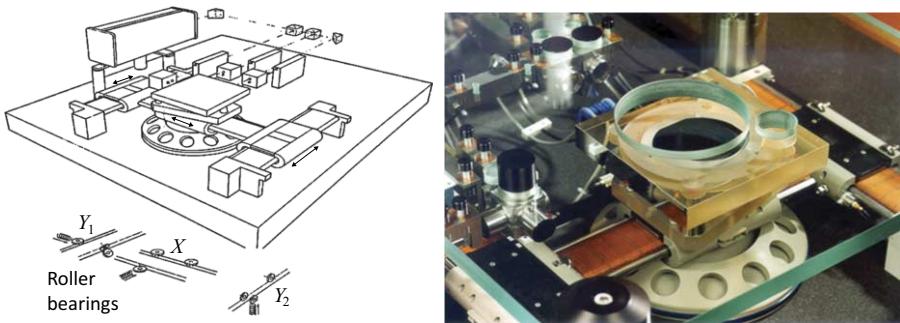
moving yoke that moves above two coils that are enlarged in the step (5) to give a larger range.

Unfortunately the large coil will induce a reluctance force acting on the moving part of the yoke (6), driving it in the outer positions where the reluctance for the magnetic field of the coil is the least. This problem is largely reduced by choosing shorter coils and applying electronic commutation as shown in step (7).

The windings around the ferromagnetic yoke consist of two coils that are interlaced with an alternating winding direction. This means that the first section of coil one is wound clockwise, followed by a clockwise wound first section of coil two, again followed by the counterclockwise wound second section of coil one, and so on.

With commutation , reversing the current direction at any coil, the sections of that coil that are wound clockwise give a counter clockwise current and the other way around. This is shown in steps (8 – 10) for a movement to the right. In principle the stator with the coils and the inner ferromagnetic yoke can be made infinitely long by adding more sections to each coil and in step eleven also the moving part is shown to be multiplied to increase the force.

The main problem of this configuration is the efficiency, as the current has to run through the entire coil, including those coil sections that are outside the permanent magnetic field. This disadvantage could be solved by feeding each coil section with a separate power amplifier but that was not done in the real system in order to avoid disturbing forces due to switching the coils on and off. It also would require a lot of separate wires while otherwise one actuator only needs four wires with two amplifiers to control the current in a smooth way without disturbing switching moments.



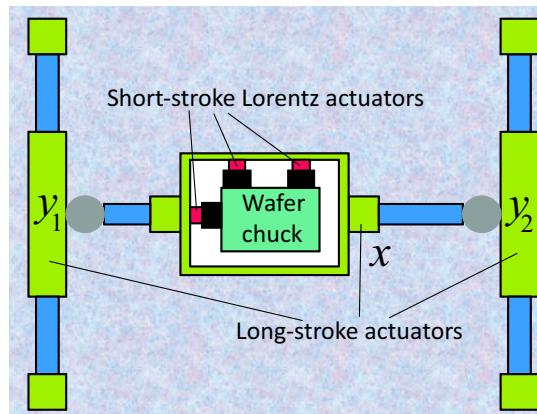
**Figure 9.12:** Original 3-D sketch and photographic image of the first wafer stage with electric linear actuators for a wafer stepper in a H-configuration. The magnetic mover is supported by roller bearings that run on rails alongside of the coil sections while an air-bearing is used to support the wafer chuck and mirror block, floating over a flat granite stone.

The pulse-width modulated design principle of the applied amplifiers was presented in Figure 6.69 of Chapter 6.

### 9.3.1.2 Multi-axis positioning

The different directions in a wafer stepper and scanner are defined in a metrological coordinate system that is oriented relative to the scanning direction and the projection lens. The horizontal direction  $x$  equals the stepping direction of the waferstage while  $y$  equals the scanning direction in both wafer and reticle stage. The vertical direction  $z$  is directed according to the optical axis of the projection lens while the rotations around the translational axes  $R_x$ ,  $R_y$  and  $R_z$  determine the remaining directions. In the first wafer stepper, the described extended-range linear Lorentz actuator only needed to provide movements in three directions in the horizontal plane as the wafer was oriented parallel to the image on the wafer chuck by mechanical means.

The standard method at that time for multi-axis positioning was to stack different single-axis motion systems on top of each other like shown in Figure 9.11. A three-axis stage, working in the  $x - y - R_z$  direction, generally consisted of two linear drives with a rotation table on top. Due to the stacking, the driving force could never be directed towards the centre of mass of the moving system. The first actuator had to drive also the other actuators while only the last actuator was close to the centre of mass of the moving stage.



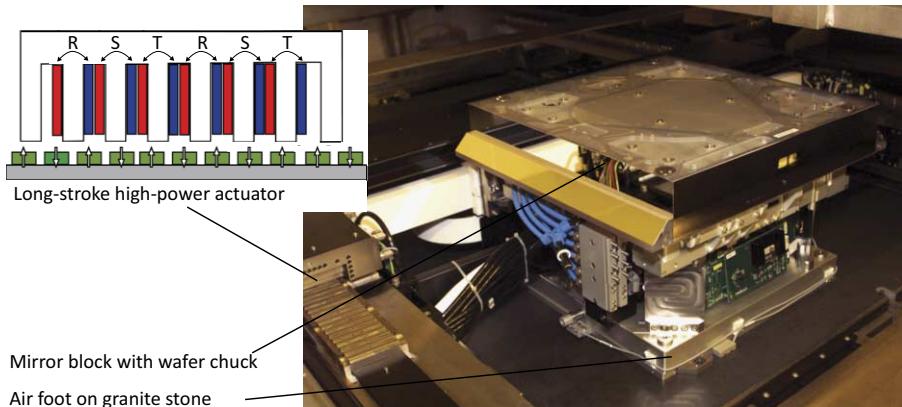
**Figure 9.13:** Schematic presentation of a 3-axis long-stroke, short-stroke wafer stage in a H-configuration. Both long-stroke and short-stroke actuators work in the same plane through the centre of mass of the wafer table. The rotation along the  $z$ -axis is achieved by the difference in positioning of the  $y$  actuators.

For this reason a different configuration was chosen, named after the letter “H” of the alphabet with one linear actuator for the linear movement  $x$  and two linear actuators  $y_1$  and  $y_2$  for the linear movement in the  $y$  direction, allowing a limited rotation  $R_z$  by the difference of  $y_1$  and  $y_2$ . Figure 9.12 shows this configuration where the bearings consisted both of roller bearings to support the permanent magnetic movers on the coil stator and the air foot, an air bearing floating on a flat granite stone surface to support the wafer stage itself. With later versions of the wafer stepper many improvements were added to this basic system with for instance three separate actuators to position the wafer in the three remaining directions,  $z - R_x - R_y$  that were needed in order to guarantee the vertical position and flatness of the wafer relative to the focal plane of the lens.

### 9.3.1.3 Long- and short-stroke actuation

The single-stroke extended-range Lorentz actuator from the previous section was sufficiently accurate to serve in a wafer stepper that only had to stand completely still at exposure.

With the wafer scanner, however, the exposure takes place during a movement at a constant velocity. Furthermore the increased requirements on resolution and overlay posed more severe requirements on precision and acceleration. This necessitated the design of a wafer stage without the in-



**Figure 9.14:** The wafer stage of a wafer scanner. The three-phase coil section with ferromagnetic yoke of the powerful long-stroke actuator runs over a permanent magnet stator. The mirror block is used to measure the position of the wafer stage with a laser interferometer.

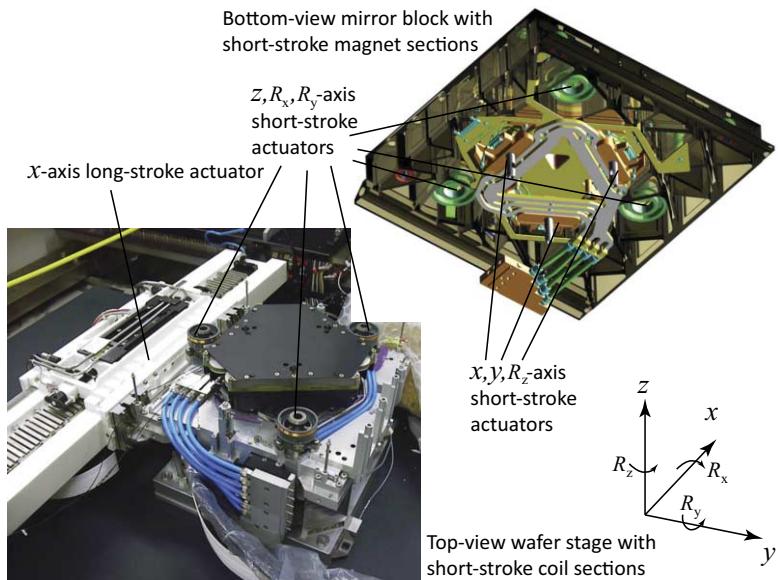
herent drawbacks of the linear actuator that was used in the wafer stepper, like the remaining parasitic forces by the commutation and the friction variations of the applied roller bearings.

The solution was to use a non-commutated *short-stroke* Lorentz actuator in cooperation with a second actuator, the *long-stroke* actuator, as shown schematically in Figure 9.13.

This long-stroke actuator is connected to the stationary part of the short-stroke actuator and only needs to position this part in such a way that the permanent magnet system of the short-stroke actuator remains in the  $\pm 0.1$  mm range around the centre of its range with the least stiffness value. This configuration allows the long-stroke actuator to operate at a reduced precision of only this same level of  $\pm 0.1$  mm, even though the precision of the short-stroke actuator needs to be on sub-nanometre levels.

The movements and vibrations of the long-stroke actuator are hardly transferred to the sensitive wafer position because of the low stiffness of the short-stroke actuator. For that reason it is allowed to use strong electronically commutated three phase actuators with ferromagnetic yokes in the long-stroke actuation system as shown in Figure 9.14.

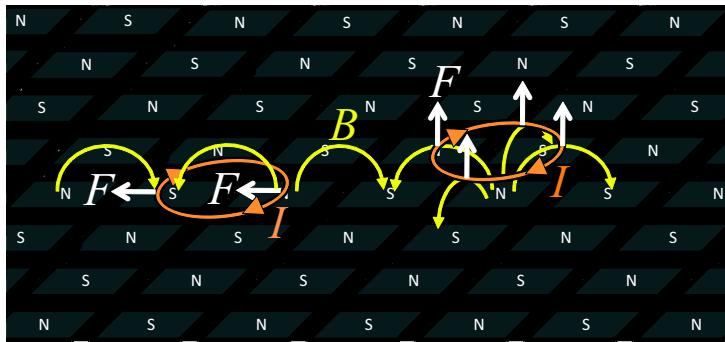
This actuator is optimised to force rather than to precision positioning and zero-stiffness with a ferromagnetic yoke around the coils. This yoke will cause *cogging* forces towards those positions where the reluctance for the permanent magnet field is smallest. Cogging forces are perceived as bumps when moving the mover by hand over the total stroke.



**Figure 9.15:** The short-stroke actuators are integrated within the mirror block. The actuation in the horizontal plane is done by flat Lorentz actuators while the out-of-plane movements are provided by loudspeaker type Lorentz actuators without a ferromagnetic yoke to prevent non-linearity by reluctance forces.

The advantage of the ferromagnetic yoke around the coils is the low reluctance path for the magnetic field that increases the force to current ratio of the actuator. This actuator can easily transport the required cables, wires and water cooling tubes for both the short-stroke coils and its own coil section.

Fortunately only the in plane motions are large in a wafer scanner, so only those motions require a corresponding long-stroke actuator. The out-of-plane motions, needed to keep the surface of the wafer in the focal plane of the lens are realised by separate short-stroke actuators only, as shown in Figure 9.15. These loudspeaker type Lorentz actuators are designed without a ferromagnetic yoke in order to prevent reluctance forces as was explained in Section 5.3 of Chapter 5.

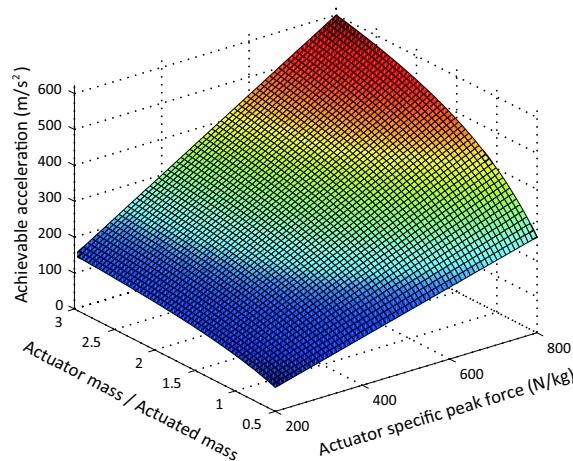


**Figure 9.16:** The forces acting on a coil, positioned in the magnetic field above an array of alternating permanent magnets. Depending on the position in the field, the forces can be in-plane, perpendicular or a combination of both.

### 9.3.2 Full magnetic levitation

The guiding of the long-stroke part of the H-configuration wafer stage is mainly realised by means of air bearings. The air foot provides the support of the heavy centre piece with the short-stroke actuator coils and is preloaded by a closed area with vacuum between the air foot and the granite stone to achieve a sufficient stiffness. The long-stroke actuators themselves have integrated air bearings that are preloaded by the attracting magnetic forces between the mover and the stator.

In future wafer scanners with EUV light of 13 nm, air is not allowed in the light path because it absorbs the light. For that reason these machines are operated in vacuum. Although it is in principle possible to realise air bearings in a vacuum chamber, by adding special measures to locally extract the air that escapes from the side of the air bearings, it is better to avoid them. For that reason a long-stroke stage has been developed that uses an electromagnetic planar actuator, working in six degrees of freedom. Its working principle is shown in Figure 9.16. A permanent magnetic platform, consisting of a plurality of alternating permanent magnets, creates a magnetic field with different field directions. A current-carrying coil will experience a force of which the direction depends on its position on the magnetic table and the direction of the current. By using different coils at different locations a configuration can be created that is able to exert forces and torques in all six degrees of freedom at any position of a free moving stage, when combined with a suitable electronic amplifier system.



**Figure 9.17:** All electric actuators have a limitation in the amount of force per unit of moving mass of the actuator itself, the specific peak force. With a given mass that has to be actuated, an increase of the acceleration would require an increase of the mass of the actuator. The relative added acceleration flattens out at a high ratio between the mass of the actuator and the actuated mass with an asymptote at the maximum specific peak force.

### 9.3.3 Limits in acceleration of reticle stage

The continuous increase of the acceleration levels of reticle stages poses several challenges that are increasingly not trivial. The main issue is that both the long-stroke and short-stroke actuator are essentially placed in series, which means that they both need to deliver the same force. This high force is especially problematic for the Lorentz actuator.

Like any other actuator also a Lorentz actuator has thermally determined limitations in its force and power capability. The high power dissipation in the resistance of the windings by the motor current must be removed by means of a closed water circuit. At a certain level of heat per unit of volume of the windings, the temperature of the windings will reach the maximum value that the insulation can sustain, determined by the heat conductivity of the windings and the cooling circuit. The volume of the windings is proportional to the size of the moving permanent magnetic part of the actuator which leads to a maximum in the amount of force per unit of moving mass of the actuator itself. This *specific peak force* of a Lorentz actuator is approximately 650 N/kg and corresponds with the maximum acceleration that the actuator could realise without the mass of the actuated

body.

A direct consequence of this relation is that the total moving mass of the stage becomes noticeable larger than only the mass of the actuated body that carries the wafer. For the required maximum acceleration of  $120 \text{ m/s}^2$  as mentioned in Table 9.1, the mass of the actuator is still below 50 % of the mass of the actuated body and this can as yet be provided without much problems, but when more acceleration is needed in the future, the total mass will rapidly increase. For this reason, present developments on the stages focus on the reduction of the moving mass by the application of hollow structures with thin plates and reinforced plastics. Unfortunately this mass reduction does not only have benefits as the sensitivity for external disturbing forces by transmission is increased. The counterbalancing effects of the mass with on the one hand the need to decrease the mass in order to increase the maximum acceleration and on the other hand the need to increase the mass in order to reduce the sensitivity for external disturbing forces is called the *mass dilemma* of precision positioning systems. In other words: "A reduced mass requires improved system dynamics that enable a higher control bandwidth to compensate for the increased sensitivity for external vibrations".

## 9.4 Position measurement

Several position measurement systems are applied in wafer scanners. The following list mentions the most important ones:

- The alignment system that measures the position of the wafer relative to the wafer stage with sub-nanometre accuracy.
- The level sensors that measures the surface profile of the wafer relative to the wafer stage with nanometre accuracy.
- Long-range laser interferometers or encoders for real-time incremental position measurement of the stages relative to the metrology frame with sub-nanometre accuracy.
- Capacitive sensors for small relative measuring distances and ranges like the vertical position of the reticle stage relative to the metrology frame with nanometre accuracy.
- Optical proximity detectors for surface measurement of the wafer with nanometre accuracy.
- Conventional encoders for less critical measurements like the internal position measurement of the long-stroke actuators with sub-micrometre accuracy.
- State-of-the-art encoders for extreme-precision measurements of very small displacements with an accuracy of several picometres that are used to measure the relative position of optical elements inside the projection lens.

The most important and complicated measurement in a wafer scanner is related to the overlay, the position of the previously exposed layers on the wafer relative to the image of the reticle. The complication in this metrology loop from image to previous layer is based on the fact that this measurement consists of several relative measurements of which some are not possible in real-time. The main cause for this problem is that the projection lens closely covers the part of the wafer that is exposed, thereby preventing any direct measurement of the image and the previous layer at that location. The only alternative to a direct measurement is an indirect measurement where the position of the previous layers on the wafer and the image location are separately measured relative to a sufficiently stable reference with a well-known position in the wafer scanner. This reference is the real-time

measured wafer stage. The 6-axis incremental position measurement of the stages is continuously available and defines the connection between the stages and the metrology frame, the central reference that is connected to the lens. This real-time measurement system is also used for the closed-loop feedback control and synchronisation of the reticle stage and wafer stage with strict requirements on latency.

A consequence of this indirect measurement is the addition of all measurement errors that occur during the different measurement steps. This addition can be calculated by means of the “root of the sum of squares” of the errors, as explained in the Chapter 8 on measuring, but it requires that the accuracy of all measurements have to remain well better than the overlay specification, in practice below one nanometre.

Fortunately the requirements on overlay of a few nanometres are only strictly relative. There is no traceable relation with the standard metre other than the order of magnitude. Traceability with a proven uncertainty level over a long period of time is not important. All calibrations are related to the previous layers on the wafer and only need to remain stable between the moment that the previous layer is measured relative to the wafer stage and that the exposure is finished. Furthermore underlining the relative nature of the measurements is the fact that the wafer itself is not stable as during the chemical processing at high temperatures the previously defined structures can be moved by deformation of the wafer. By measuring these deformations both the magnification of the lens can be adapted and the positioning of the stages can be made to accommodate to these deformations. The following sequence of different steps describes more in detail the total measurement process from the location of the previous layer to the image position:

1. The wafer is pre-aligned in the wafer handler with optical sensors to measure the location of the flat edge or the notch at one location on the side of the wafer that determines the basic orientation of the wafer.
2. A robot brings the wafer from the pre-alignment chuck to the wafer stage. The position of the vacuum gripper of the robot is alternately referenced to the pre-alignment chuck and the waferstage by a kinematically determined mechanical connection. The total resulting positioning error of the wafer on the waferstage is approximately ten micrometres.
3. The position of the previous layers on the wafer is measured by means of alignment marks that are defined during the exposure of the first

layer. The number of these reference marks can range from two till the total number of dies. Similar alignment marks, called *fiducials*, are also stably attached to the wafer stage, close to the wafer. The alignment system measures the position of the alignment marks on the wafer relative to the position of the fiducials.

4. The waferstage with the aligned wafer moves to the projection lens under real-time measurement relative to the metrology frame. The position and focal plane of the image is measured relative to the wafer stage with a special image sensor. It is assumed that this information remains stable during one or several exposure cycles.
5. The fiducials of the wafer stage are measured either to a reference inside the lens or relative to alignment marks on the reticle through the projection lens.

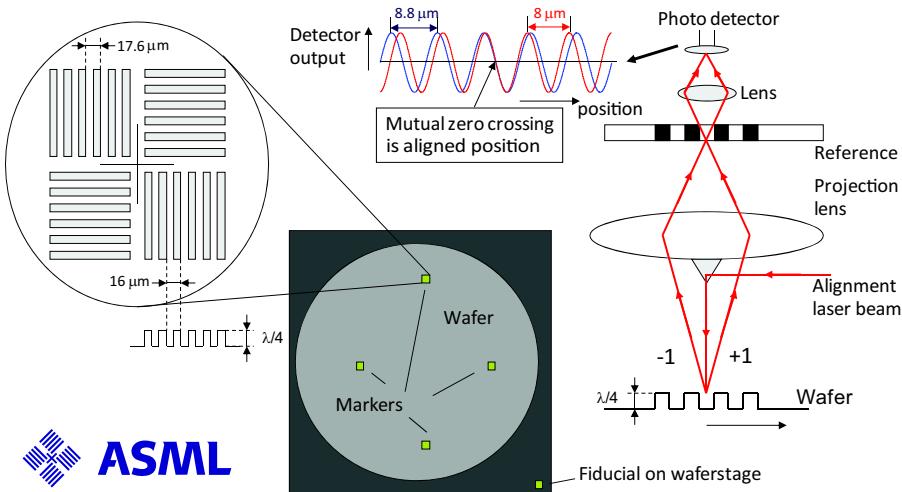
In the following subsections the alignment sensing principle, the measurement of the focal plane and the real-time metrology loop with the long-range incremental measurement systems will be presented.

#### 9.4.0.1 The alignment sensor

The principle of alignment with the alignment marks as reference for the pattern on the wafer is shown in Figure 9.18. The alignment marks are phase gratings and the measurement is comparable with the interferometric encoder of Section 8.8.1.3.

The grating is illuminated with a laser and the phase steps with a height of  $0.25\lambda$  create diffraction orders mainly in the two first orders. Both orders are recombined at the photo detector after interfering with a second reference grating. The interference irradiance depends on the phase relationship of these orders. As was explained with the interferometric encoder, the phase shift by the movement in the  $+1^{\text{st}}$  order is opposite and equal in magnitude to the phase shift of the  $-1^{\text{st}}$  order, resulting in a doubling of the combined differential phase shift, thereby increasing the measurement resolution with a factor two.

The alignment is done by a scanning movement of the wafer stage through the range where optimal alignment should take place and the zero crossings of the irradiance signal are measured. These zero crossings repeat with every half grating period due to the doubling effect of the two diffraction orders which limits the range of this alignment system to half a period of



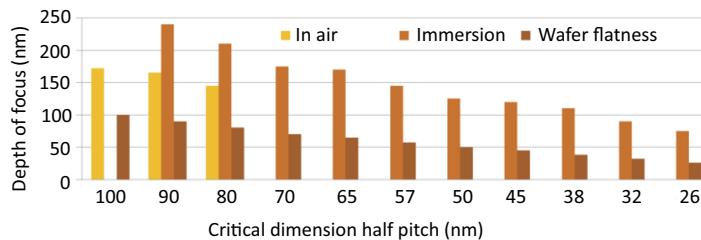
**Figure 9.18:** The alignment marks are reflective phase gratings with different periodicity, to increase the capture range. One way to measure the markers is shown at the right, where both first orders are imaged on a reference transparent phase grating and their disturbance is detected by a sensor. This configuration increases the position sensitivity with a factor two. Note the alignment mark in the ASML logo.

the alignment grating, being  $8 \mu\text{m}$ .

This value is too small for the wafer handler to reliably position the wafer inside the capture range of the alignment system. A misalignment of  $8 \mu\text{m}$  would mean that the entire exposure becomes erroneous.

To avoid this problem and increase the capture range, a Vernier principle is applied by simultaneously measuring a second grating with a 10 % different grating period. The zero crossing of the irradiance by the two gratings will only coincide in one on every ten periods of the irradiance signal of one grating, increasing the range with a factor ten. In practice a capture range of approximately  $40 \mu\text{m}$  is adhered in order to avoid errors by outliers in the measurement.

An interesting effect of the phase grating is that the phase shift by the displacement remains the same irrespective of the height of the grating. As was shown with the interferometric encoder, the phase shift is only determined by the ratio between the displacement and the grating period. This implies that the position detection is not affected by the height of the grating. On the other hand, the irradiance in both 1<sup>st</sup> orders does depend on the height of the grating with a maximum at a step height of  $0.25 \cdot \lambda$ . At



**Figure 9.19:** The required depth of focus (DOF) as function of the critical dimension. After the introduction of immersion, the requirements were less stringent.

a certain moment Chemical Mechanical Polishing (CMP) of the metal layers was introduced in IC manufacturing to flatten the surface in order to achieve a better focusing performance and it was feared at the time that this CMP-process would erase the alignment marks. Fortunately it appeared that a much shallower grating than the optimal value still provides just sufficient signal for alignment. With some additional measures on sensitivity of the optical sensors the remaining phase differences over the flattened gratings are still detectable with sufficient accuracy.

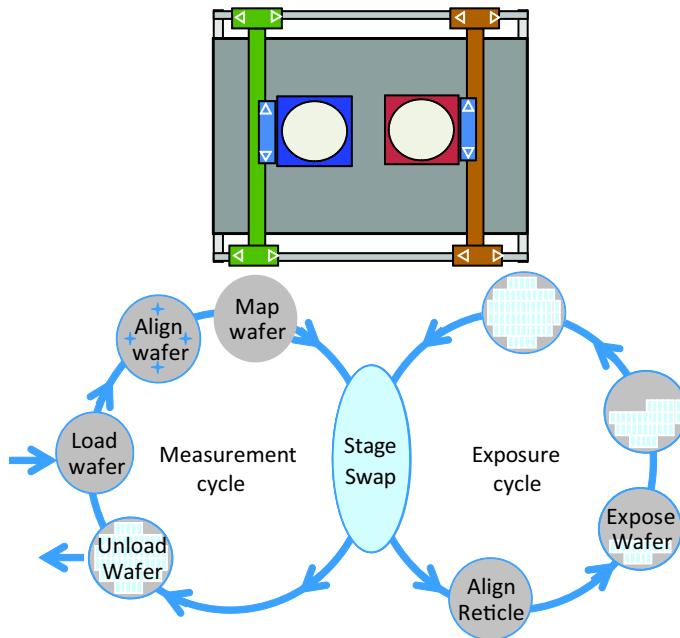
#### 9.4.1 Keeping the wafer in focus

The high numerical aperture that is necessary for a better resolution automatically implies a decreased depth of focus (DOF) as was explained in Chapter 7. The three dimensional structures on the surface of a wafer, that are created during the many manufacturing process steps, have made it increasingly difficult to guarantee a sufficiently flat surface to accommodate this continuously decreasing depth of focus of the projection lens. Figure 9.19 shows the relation between the depth of focus and the critical dimension. A focal error of less than 100 nm over a die of 25 mm is a challenging requirement. To achieve such extreme specifications, this problem is approached from two sides. The first step is to guarantee a certain flatness of the wafer. As previously explained the deposited metal layers for the wiring are often treated by a CMP-process to keep the surface non-planarity locally within acceptable levels to less than 50 nm over a die.

The second step to achieve a sufficiently small focus error is to actively position the wafer within the focal area at the region of the exposed die. This process is also called the *levelling* of the wafer and with the first generations of the wafer scanner the levelling was done by means of real-time

measurement of the wafer surface ahead of the lens during scanning. This measurement was done by means of an optical triangulation method like explained in Section 8.6.1.2 of the previous chapter on measurement. A servo tracking feedback control system kept the surface approximately parallel to the image in the focal plane. Unfortunately feedback control can never prevent some remaining focal error due to the positioning system dynamics and the limited control gain. A better solution was found to determine the height profile of the wafer prior to exposure and use that information in feedforward control at exposure.

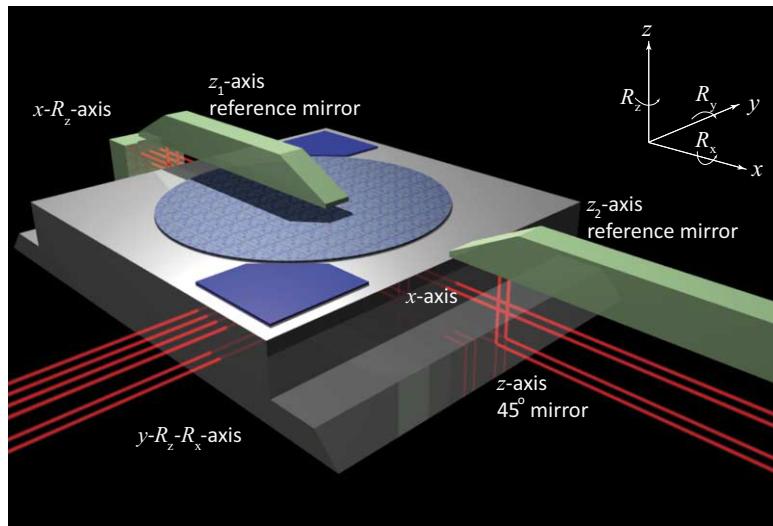
With a single wafer stage a separate measurement of the surface topology of the wafer would require a decrease in the productivity due to the additional time for the measurement. The dual wafer stage with parallel measurement and expose cycles was designed to solve that problem.



**Figure 9.20:** The process steps in a dual-stage wafer scanner between loading and unloading consist of two cycles. In the measurement cycle the wafer is aligned to the wafer stage and the surface is mapped for focus. In the exposure cycle the wafer stage is aligned to the reticle and the wafer is exposed.

### 9.4.2 Dual-stage measurement and exposure

Productivity has been the main driver for the decision to design a wafer scanner with parallel processing of the measurement and the exposure cycle. Next to the possibility to precisely measure all dimensional details of a specific wafer without sacrificing throughput, the increased utilisation of the expensive lens is another important reason for this duplication of the wafer stage. With a dual wafer stage the lens can expose one wafer during the time that the other wafer is measured and as a consequence the lens is continuously utilised. Figure 9.20 shows the process steps of a dual-stage wafer stage. The measurement cycle of one wafer is executed synchronous with the exposure cycle of another wafer. As a result of the measurement, the exposure starts with a wafer that is fully known in respect to the position of the alignment markers to the wafer stage and the height profile of the surface. This enables a maximum use of feedforward control.



**Figure 9.21:** A mirror block for 6-axis position measurement by means of laser interferometry. The  $z$ -axis is realised with a mirror under  $45^\circ$  that reflects the light upwards to a reference mirror. All axes are dual path plane mirror interferometers and  $R_y$  is determined from two  $z$  interferometers.

### 9.4.3 Long-range incremental measurement system

The long-range measurement systems in the wafer- and reticle stage have to be able to track the position with the specified accuracy and velocity levels without adding latency which would reduce the accuracy of the synchronisation between the wafer stage and reticle stage.

Until the end of the last century only laser interferometer measurement systems were capable of measuring large distances with nanometre level resolution and a latency timing that neither limited the bandwidth of the feedback control nor interfered with the synchronisation of the stages. Sample rates far above 20 kHz are presently state of the art and with suitable synchronised phase detection electronics a large number of measurement axes could be realised that all worked in parallel.

Figure 9.21 shows the mirror block of a wafer stage, showing the laser beams that are used to realise a six-axis real-time measurement system.

A special configuration is used to derive the  $z$ -axis measurement signal with a  $45^\circ$  mirror to direct the light upwards from the wafer stage. The light is reflected back by a stationary reference mirror that is connected to the metrology frame and returns via the  $45^\circ$  mirror to the interferometer.

With this configuration, the total optical path length of this beam is determined both by the  $x$  and the  $z$  distance. By subtracting the directly measured incremental  $x$  movement from another laser beam, the  $z$  position is obtained.

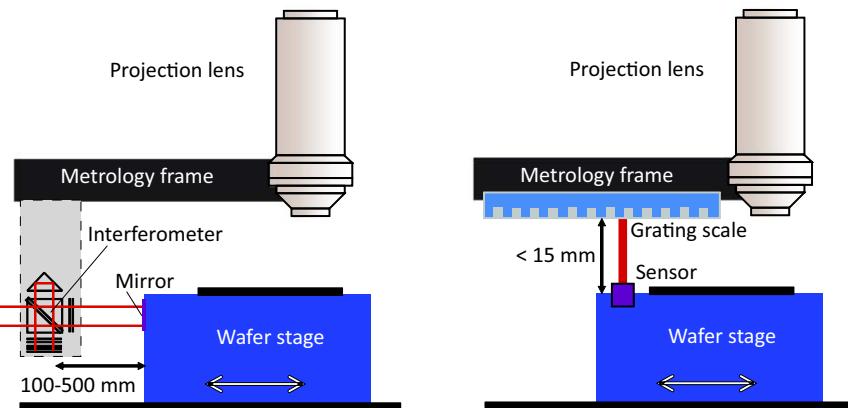
The accuracy of laser interferometer measurement systems is strongly influenced by the refractive index of air according to the Edlén equation (8.86) that was presented in Section 8.8.2.3 of the previous chapter on measurement. The main problem with interferometer distance measurement systems is that they measure over a long distance through air. At large distances, small variations in the light velocity in air, that are induced by the temperature and/or the pressure, can have a significant influence on the measurement. To keep the accuracy of the measurement on acceptable levels, *air showers* are used to condition the air over the full optical measurement path. The temperature of the air is required to be controlled within about 1 mK stability over an exposure cycle. Furthermore the air has to flow at a higher speed than the maximum speed of the stages in order to prevent mixing of the conditioned air from the air showers with air coming from elsewhere at a different temperature. These requirements have resulted in significant developments that always have just met the required specifications. Still it is assumed that on the long run this method will no longer be adequate and that also the high speed of the conditioned air might excite dynamic resonances in the sensitive parts.

Fortunately another solution is found in a planar reflective-encoder system with a phase grating that measures in all six directions with a sufficient resolution to be applied as the long-range measurement system of a wafer scanner. This alternative has two advantages over the laser interferometer. First of all the thermally induced refractive index changes have far less effect because of the short path through air of less than 15 mm. The thermal sensitivity is reduced with a factor 100 to  $\Delta L/L = 2 \cdot 10^{-8}$  per degree Kelvin.

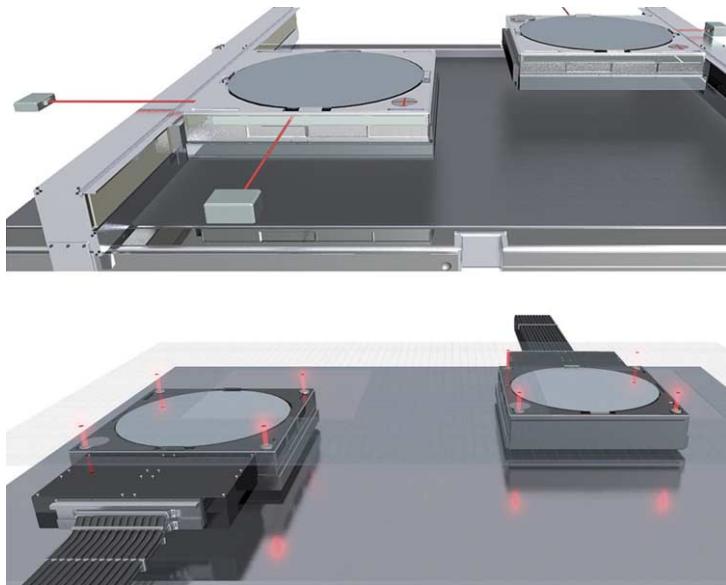
The second advantage has to do with the stability of the real-time metrology loop as is explained with the help of Figure 9.22.

#### 9.4.3.1 Real-time metrology loop

The real-time metrology loop is a subset of the full metrology loop that was defined at the start of this section and contained also measurements that were not performed in real-time. The real-time metrology loop assumes that all alignment steps are done well and remain stable during the exposure. This means that the real-time metrology loop connects the lens via the metrology frame to the wafer on the waferstage.



**Figure 9.22:** Two methods for long-range incremental position measurement. The laser interferometer, shown at the left has long been the standard but is partly replaced by the plane encoder system as shown at the right. The difference is seen in the length and stability of the metrology loop and a reduction of thermally induced instability by refractive index changes of air.



**Figure 9.23:** The combination of the H-configuration with the laser interferometer measurement system on the top and the plane encoder with the planar actuation system at the bottom clearly illustrate the differences.

The first measurement in this real-time metrology loop is the position of the lens with respect to the metrology frame. As was shown previously the connection between the lens and the metrology frame is not stiff and can even contain active dampers to control vibrations. Because of the low stiffness, separate optical or capacitive sensors are applied to measure any small displacements of the lens relative to the metrology frame.

The second measurement in the real-time metrology loop is the position of the wafer stage with respect to the metrology frame by the long-range measurement system. With laser interferometers, the interferometer part is firmly connected to the metrology frame at a distance from the optical axis that is determined by the maximum movement range of the wafer stage. This distance requires the metrology frame to be stable over the total distance from the lens to the interferometer. In principle this problem can be solved by adding a reference measurement axis from the interferometer to the lens but that would increase the induced errors by the refractive index variations of air due to this additional measurement.

With the planar encoder, the measurement grating, called the *grid plate* is stationary mounted on the metrology frame and consists of thermally inert material like Zerodur, a glass ceramic made by Schott with zero thermal expansion in a limited temperature range. This approach both reduces the length of the metrology loop through air and decreases the stability requirements of the metrology frame.

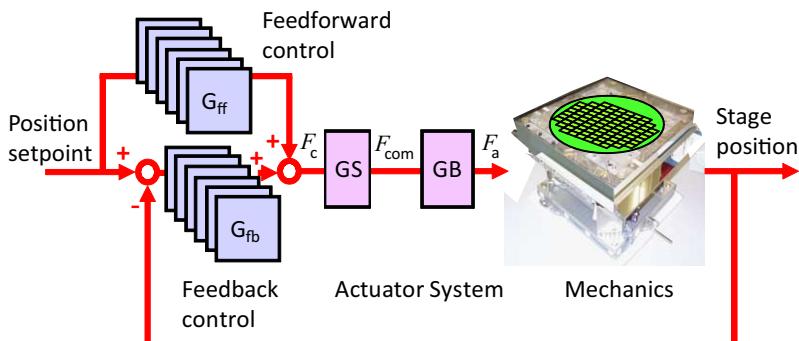
The last part of the metrology loop within the wafer stage from the measurement mirror to the wafer is quite simple and straightforward for the laser interferometer option. The mirror block forms a “monolithic” structure with the wafer chuck that holds the wafer firmly attached on the *wafer table* by vacuum and this monolithic structure is also made from a thermally inert material like Zerodur.

The plane encoder version has a more complicated last part of the loop to the wafer as the sensor part is located on the moving part of the wafer stage. The potential problems regarding stability of this part are solved by both fully integrating all optical parts in one unit and by continuous calibration with redundant measurement axes. In principle four sensors are used that each determine more degrees of freedom. The superfluous degrees of freedom are used to compare and calibrate all sensors at every exposure cycle. Figure 9.23 shows the different drive and measurement configurations for comparison.

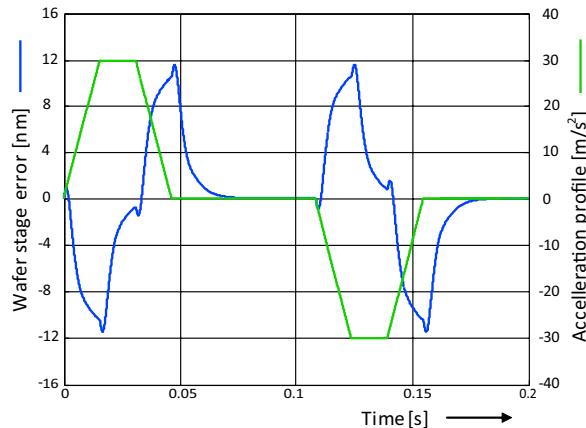
## 9.5 Motion control

Figure 9.24 shows the basic principle of the motion control of the wafer stage. A modern control expert will miss elements like Multiple Input Multiple Output (MIMO) controllers, observers and other more recently developed control approaches.

It is true that the total controller itself consists only of six separate Single Input Single Output (SISO) controllers for all directions with feedforward and feedback paths. The feedback controller uses a well tuned classical Proportional, Differentiating and Integrating PID-control algorithm. The six-axis machine coordinate system for the position measurement and the controller has its orientation at the centre of the image position. This is not the place where the centre of mass of the wafer stage, nor the actuators of the wafer stage are located. For that reason the six control forces  $F_c$  are first transformed by a *Gain Scheduling* matrix operation into the forces  $F_{com}$  that should be exerted at the centre of mass of the positioning stages. The gain scheduling matrix is constantly adapted to the actual position of the stage. A second transformation is applied to determine the corresponding forces of the actuators  $F_a$  by means of a *Gain Balancing* matrix operation. The gain balancing matrix has to be well tuned to the physical hardware of the wafer stage in order to compensate for performance deviations due to tolerances in the different functional parts.



**Figure 9.24:** The position control of the wafer stage is a SISO 6-axis PID-control system. The Gain Scheduling matrix (GS) transforms the control forces ( $F_c$ ) in six degrees of freedom, according to the machine coordinate system into the forces ( $F_{com}$ ) of the wafer stage coordinate system around the centre of mass. The Gain Balancing matrix (GB) transforms these forces into the actuator forces ( $F_a$ ).



**Figure 9.25:** The wafer stage error as function of time induced by a mismatch of 0.1% of the feedforward force when a well-tuned feedback is applied.

The reason for this rather simple and straightforward approach in control is twofold. First of all this methodology allows a more direct investigation of observed errors in the system. This shortens the time for trouble shooting at the customer site, when the system has failed, or when parts have been exchanged and the system should be tuned again. The second reason is that the feedback part only plays a role in the robustness of the control to deviations of the system parameters from the ideally modelled situation as will be further explained in the following section.

### 9.5.1 Feedforward and feedback control

A property of feedback control is the fact that it needs an error to act anyway, so it is allowed to say that **Feedback is always (too) late**. The related delay manifests itself in the *settle time* that is the time, needed to reduce the effects of an disturbing event to the specified maximum error.

Feedback should only be applied to correct errors that could not be avoided by feedforward control due to random disturbances or uncertainties in the properties of different elements of the system. In the design of a wafer scanner the focus is on creating a dynamic system that behaves as deterministic as possible, mainly because of the impact of settle time on throughput.

A small example can place this requirement in a realistic perspective. Imagine a wafer stage with the specifications as mentioned in the list of Table 9.1, a mass of 20 kg, accelerating with  $30 \text{ m/s}^2$  to a maximum speed of 0.5 m/s. With these numbers the required force during acceleration becomes

600 N and the acceleration time takes  $\approx 17$  ms.

Assume that this situation is the result of a pure open-loop inertia-based feedforward controlled system with a maximum error in the force of only 0.6 N ( $=0.1\%$ !). Then a small calculation gives the following position error after the 17 ms acceleration time.

$$\varepsilon = 0.5at^2 = 0.5 \cdot \frac{0.6}{20} (17 \cdot 10^{-3})^2 \approx 4 \cdot 10^{-6} \text{ m} \quad (9.7)$$

This is still a factor 400 above the required 10 nm level. Improving the feedforward force to the required level would mean a total error in the order of 1 ppm ( $1 : 10^6$ ) and those are levels that can only be realised in electronics. This all clearly indicates the need for an additional feedback control action to reduce the remaining error and Figure 9.25 shows a realistic graph of the modelled performance of a waferstage with a well-tuned PID feedback controller that illustrates the strong improvement that is achieved. The feedback control starts almost immediately after the beginning of the acceleration profile and keeps the overall error at just more than 11 nm which almost meets the specifications.

This proof of the value of feedback should however not be seen as a reason to reduce the requirements for the feedforward control. When the accuracy of the feedforward control action would for instance be reduced with a factor three that the graph shows that the settle time to reduce the corresponding higher peak error with a factor three is almost 10 ms.

This example reduction of the feedforward accuracy with a factor three has the following consequences for the throughput of the machine. An average 300 mm wafer has 100 dies so 100 times acceleration and deceleration. These 10 ms settle time per scan would mean approximately 1 second loss from the 20 seconds total exposure time of the machine. This 5 % loss reduces directly 5 % of the customer value of the wafer scanner!

When the requirements in future become more strict there are in principle only two ways to reduce the settle time. The first is to increase the control bandwidth frequency. This however requires further improvements on the mechanical dynamics that were already assumed to be optimal and it is not expected that this can be easily improved without increasing cost. The only other solution that remains is an improved accuracy of the feedforward control system. The errors in the feedforward model are mainly determined by changing properties of the actuators due to temperature and in practice such a reduced error level can only be maintained by continuously calibrating the total system.

### 9.5.2 The mass dilemma

The above example did not include errors caused by external disturbances. With all measures taken to its maximum, in practice still the last errors have to be corrected in less than a just allowable 10 ms settle time and there the mass dilemma returns in its full magnitude!

A reduced moving mass is preferred in order to reduce the electrical power that is necessary for the required high levels of acceleration. A reduced mass will also limit the reaction forces and will indirectly lead to reduced errors by nonlinearities and dynamic disturbances in the actuators.

Unfortunately a reduced mass will proportionally increase the sensitivity for external disturbance forces through elastic and damping elements. This means that an increased feedback control gain would be needed to correct for these errors with a corresponding higher bandwidth frequency. One might think that a low mass construction will show higher eigenfrequencies that enable to achieve this increased bandwidth. Unfortunately this is not a fully correct assumption as a reduced mass without reduced overall dimensions can only be realised by a combination of thin structures. Those structures are not always sufficiently stiff in all directions to guarantee the required high eigenfrequencies.

Research is done at several places to compensate these potentially harmful eigenmodes by means of *overactuation*. Overactuation implies the use of more actuators than the number of coordinate directions that have to be controlled. By optimising the placement of the actuators in respect to the mode shapes of interest, the excitation of the lower-frequency eigenmodes can be avoided. As soon as these techniques are really applied, the simple control configuration of Figure 9.24 is no longer possible as the actuation axes are dynamically well-coupled giving a space matrix with many numbers beside the diagonal.

## 9.6 Main design rules for precision

Next to the purpose of pin-pointing the important value of mechatronics for a major technological development of this era, this chapter was written to emphasise some important universal design rules. Following these rules will help to achieve a high precision when designing high-speed production systems like wafer scanners:

1. **Know your problem.** Only start designing when the requirements and specifications are clear.
2. **Don't create your own problem.** Potential disturbance sources should be solved at the source.
3. **Know your hardware.** The first design goal is to define an extremely well-balanced hardware concept that offers the best possible deterministic dynamic behaviour of all systems combined. Mechatronics is an integral approach on design.
4. **Do not think that software solves your problem.** Changes in hardware are very expensive and time consuming. In most cases software is only capable to camouflage flaws in the hardware by introducing longer waiting times.
5. **Feedback only for robustness.** Make maximum use of feedforward control and use feedback only for robustness against real undetermined error sources.
6. **Measuring is knowing.** Anything can only be precise when it can be determined that it is precise. This means that all aspects of the metrology loop need to be fully mastered.
7. **Timing is essential.** Be aware of the timing of all elements in the control loop. The latency in digital measurement systems can be detrimental for a stable feedback loop.

Especially the rule on reliable knowledge of the hardware in relation to software is of paramount importance. There is a growing belief especially with younger people that software will ultimately solve all our problems. It is indeed true that, although virtually absent in this book, software has proved to be crucial for the integration and control of complex high-tech equipment. The flexibility of software in combination with the extreme high computing speed of computers has tremendous power. It should, however,

never be forgotten that a bad mechanical design can never be mended by software. A loose bolt will stay loose as no force can be transferred by software.

Of course far more subtle aspects play a role in this perception but it is safe to state that the ultimate of performance of a complex mechatronic system will always require the ultimate performance from all parts simultaneously. In this chain of performance the hardware elements will remain the limiting factor for the maximum attainable performance of most high performance mechatronic systems due to their intrinsic complexity and physical constraints.

# Appendix

## Recommended other books

The following books are recommended for further reading, when more in depth knowledge of the different disciplines within mechatronics is required.

The first two books are a useful addition to the theory on **mechanics of mechatronic systems**, because they present mechanisms, transmissions, bearings and several other aspects of mechanical engineering that are not presented in this book. The first book is based on the practical experience within the Engineering departments of Philips Electronics on the dynamics of mechanisms. Earlier versions were written in Dutch, the original by professor Wim van der Hoek and refined by professor Rien Koster, both from the Eindhoven University of Technology. Recently a fully re-written version is published in English by professor Herman Soemers from the University of Twente:

### **Design principles for precision mechanisms**

H.M.J.R.Soemers

Publisher: T-Point Print VoF (2010)

ISBN: 978-9036531030

The second book on mechanical design is written by Anton van Beek from Delft University of technology with an emphasis on reliability, friction phenomena and bearings:

### **Advanced Engineering design – lifetime performance and reliability**

Anton van Beek

Publisher: Tribos (2009)

ISBN-13: 978-9081040617

Regarding **motion control**, the following four books provide ample information that directly connects with the described methodology in Chapter 4, including the relations with the physical aspects of the controlled plant:

**Feedback Systems: An Introduction for Scientists and Engineers**

Karl Johan Åström, Richard M. Murray

Publisher: Princeton Univ Press (2008)

ISBN-13: 978-0691135762

**Digital Control of Dynamic Systems**

Gene Franklin, J. David Powell, Michael L. Workman

Publisher: Addison Wesley (1998)

ISBN-13: 978-0201820546

**Computer-Controlled Systems: Theory and Design**

Karl Johan Åström, Björn Wittenmark

Publisher: Prentice Hall (1996)

ISBN-13: 978-0133148992

**Advanced PID Control**

Karl Johan Åström, Tore Hägglund

Publisher: Isa (2005)

ISBN-13: 978-1556179426

For **optics** the following two books are recognised as being the prime source of reference on this subject:

**Principles of Optics:**

Electromagnetic Theory of Propagation, Interference and Diffraction of light

Max Born, Emil Wolf

Cambridge University Press (1999)

ISBN-13: 978-0521639217

**Optics**

Eugene Hecht

Publisher: Addison Wesley (2002)

ISBN-13: 978-0805385663

On **physics in general**, including light and waves, the many books from Richard Feynman are all worthwhile reading. The following two examples give a nice overview of his ability to teach the difficult material in clear manner, not only in the classroom where his style of lecturing became famous, but also by the examples and understandable structure of the material.

The first and most comprehensive work is based on the “Feynman lectures on physics”, that were given to undergraduate students at the California Institute of Technology. The textbook was originally written by the staff of Caltech in 1964 based on recordings of the lectures and is now available in a newly edited full set:

**The Feynman Lectures on Physics**, boxed set: The New Millennium Edition

Richard P. Feynman, Robert B. Leighton, Matthew Sands

Publisher: Basic Books (2011)

ISBN-13: 978-0465023820

The second example is a much smaller book that was written to explain the principles of the theory on quantum-electrodynamics to people from outside the particle-physics community. This very readable and even sometimes amusing booklet is referred to, when explaining photons:

**QED: The Strange Theory of Light and Matter**

Richard P. Feynman, A. Zee (intro)

Publisher: Princeton University Press (2006)

ISBN-13: 978-0691125756

For **measurement**, the following book has often been used in our lectures at the university:

**Principles of Measurement Systems**

John P. Bentley

Publisher: Prentice Hall (2005)

ISBN-13: 978-0130430281

When more knowledge is needed on the physical theory and application of **electromechanics**, dealing with all electromagnetic forces that work on objects, the following books gives these relations with examples from engineering. The first book has more emphasis on physics while the second and third book are more focused on the application in a mechatronic environment including power amplifiers.

**Continuum Electromechanics**

James R. Melcher

Publisher: The MIT Press (1981)

ISBN-13: 978-0262131650

**Electromechanical Systems and Devices**

Sergey Edward Lyshevski

Publisher: CRC Press (2008)

ISBN-13: 978-1420069723

**Electric machinery**

A.E. Fitzgerald, Charles Kingsley Jr., Stephen Umans

Publisher: McGraw-Hill (2002)

ISBN-13: 978-0073660097

For **electronics** the following recognised reference covers both analogue and digital electronic circuit design:

**The art of electronics**

Paul Horowitz, Winfried Hill

Publisher: Cambridge University Press (1989)

ISBN-13: 978-0521370950

# Nomenclature and abbreviations

## Nomenclature

$a$	[m/s <sup>2</sup> ]	Acceleration.
$A$	[ $\cdot$ ]	Amplitude.
$A$	[m <sup>2</sup> ]	Surface of Cross Section (SCS).
$A_c$	[m <sup>2</sup> ]	SCS of a coil.
$A_{c,w}$	[m <sup>2</sup> ]	SCS of the windings in a coil.
$A_g$	[m <sup>2</sup> ]	SCS of an air gap.
$A_i$	[m <sup>2</sup> ]	SCS of the flux path inside a coil.
$A_m$	[m <sup>2</sup> ]	SCS of a permanent magnet.
$A_o$	[m <sup>2</sup> ]	SCS of the flux path outside a coil.
$A_y$	[m <sup>2</sup> ]	SCS of the flux path inside a yoke.
$b$	[ $\cdot$ ]	Wien's constant.
<b>B</b>	[T]	Magnetic field.
$B$	[T]	Magnetic flux density (MFD).
$B_g$	[T]	MFD in the air gap.
$B_{g,c}$	[T]	MFD in the air gap induced by current.
$B_{g,m}$	[T]	MFD in the air gap induced by magnet.
$B_r$	[T]	Remnant MFD.
$B_s$	[T]	Saturation MFD.
$B_w$	[T]	MFD inside a coil winding.
$B_y$	[T]	MFD inside a ferromagnetic yoke.
$c$	[ $\cdot$ ]	Dimensionless constant or factor.
$c$	[m/s]	Velocity of light.
$c$	[Ns/m]	Damping coefficient.
$c_d$	[V/m]	Deformation constant piezo accelerometer.
$c_i$	[Is <sup>2</sup> /m]	Current sensitivity piezo accelerometer.
$c_s$	[ $\cdot$ ]	Sensitivity factor in Wheatstone bridge.
$c_v$	[Vs <sup>2</sup> /m]	Voltage sensitivity piezo accelerometer.
$C$	[F]	Capacitor value.
$C, C$	[m/N]	Compliance, Transfer Function (TF).
$C_{fb}(s)$	[ $\cdot$ ]	TF of a feedback controller.
$C_{pd}(s)$	[ $\cdot$ ]	TF of a PD-type feedback controller.
$C_{pid}(s)$	[ $\cdot$ ]	TF of a PID-type feedback controller.
$C_{ff}(s)$	[ $\cdot$ ]	TF of a feedforward controller.
$C_m$	[m/N]	Mass compliance.
$C_s$	[m/N]	Spring compliance.

$C_t$	[m/N]	System compliance.
$C_d$	[m/N]	Damper compliance.
$\mathbf{d}$	[m/V]	Piezoelectric coefficient matrix.
$D, \mathbf{D}$	[C/m <sup>2</sup> ][NVm]	Electric flux density or displacement.
$E, \mathbf{E}$	[V/m], [N/C]	Electric field.
$\mathbf{E}_f$	[V/m], [N/C]	Induced electric field by change of flux.
$E$	[J]	Energy.
$E_C$	[J]	Energy contained in a capacitor.
$E_L$	[J]	Energy contained in an inductor.
$f$	[Hz]	Spatial or temporal frequency.
$f$	[m]	Focal length.
$f(f), F(f)$	[-]	TF in the temporal frequency domain.
$f(s), F(s)$	[-]	TF in the Laplace domain.
$f(t), F(t)$	[-]	TF in the time domain.
$f(\omega), F(\omega)$	[-]	TF in the angular frequency domain.
$f_0$	[Hz]	Temporal natural or corner frequency.
$f_a$	[Hz]	Anti-resonance frequency of decoupling body.
$f_c$	[Hz]	Unity-gain cross-over frequency.
$f_c$	[Hz]	Centre frequency of dual laser beam.
$f_d$	[Hz]	Doppler shift frequency.
$f_d$	[Hz]	Differentiating corner frequency of D-control.
$f_i$	[Hz]	Integrating corner frequency of I-control.
$f_m$	[Hz]	Measured frequency.
$f_N$	[Hz]	Nyquist frequency limit of sampling.
$f_p$	[Hz]	Frequency of a photon.
$f_s$	[Hz]	Switching frequency of SMPA.
$f_s$	[Hz]	Split frequency of dual laser beam.
$f_t$	[Hz]	Taming corner frequency of D-control.
$F$	[N]	Force.
$\hat{F}$	[N]	Amplitude of force.
$F_a$	[N]	Force of actuator.
$F_d$	[N]	Force due to deformation or displacement.
$F_c$	[N]	Control force.
$F_{\text{com}}$	[N]	Force at centre of mass.
$F(s)$	[-]	Transfer function of a filter.
$\mathcal{F}_e$	[V]	Electromotive force.
$\mathcal{F}_m$	[A]	Magnetomotive force.
$G$	[m <sup>2</sup> sr]	Etendue.
$G$	[-]	Gain of amplifier.
$G_a$	[-]	Closed-loop amplifier TF (gain).

$G_f$	[-]	TF of operational amplifier feedback loop.
$G_o$	[-]	Open-loop TF of operational amplifier.
$G_v$	[-]	Voltage gain of transistor.
$G(s)$	[-]	Transfer function of the plant.
$G_{t,ff}(s)$	[-]	Total TF of plant with feedforward controller.
$h$	[-]	Planck's constant.
$h_c$	[m]	Winding height of a coil.
$h_{c,g}$	[m]	Overlap winding height of a coil and air gap.
$h_g$	[m]	Height of air gap in the motion direction.
$\mathbf{H}$	[A/m]	Magnetizing field.
$H$	[A/m]	Magnetic field strength.
$H_c$	[A/m]	Coercitive magnetic field strength.
$H_w$	[A/m]	Magnetic field strength inside a coil winding.
$I$	[A]	Electric current.
$I$	[W/m <sup>2</sup> ]	Intensity.
$\hat{I}$	[A]	Amplitude of alternating electric current.
$I_b$	[A]	Base current of a transistor.
$I_c$	[A]	Collector current of a transistor.
$I_d$	[A]	Diode current.
$I_e$	[A]	Emitter current of a transistor.
$I_g$	[A]	Gate current of a MOSFET.
$I_i$	[A]	Input current of an electronic circuit.
$I_l$	[A]	Current in a load.
$I_m$	[A]	Induced current by movement of actuator.
$I_N$	[A]	Norton Equivalent current.
$I_o$	[A]	Output current of an electronic circuit.
$I_r$	[W/m <sup>2</sup> ]	Irradiance.
$I_{r,c}$	[W/m <sup>2</sup> ]	Combined Irradiance.
$I_{r,i}$	[W/m <sup>2</sup> ]	Irradiance of incident radiation.
$I_{r,r}$	[W/m <sup>2</sup> ]	Irradiance of reflected radiation.
$I_s$	[A]	Electric current caused by the source.
$I_t$	[A]	Total parallel current in an electric circuit.
$\mathbf{J}$	[A/m <sup>2</sup> ]	Electric current density.
$k$	[N/m]	Stiffness.
$k_a$	[N/m]	Stiffness of anti-resonator.
$k_d$	[-]	Differentiating gain factor of D-control.
$k_d$	[-]	Differentiating gain factor of D-control.
$k_f$	[-]	Stiffness of flexure spring.
$k_i$	[-]	Integrating gain factor of I-control.
$k_n$	[N/m]	Negative stiffness of plant.

$k_p$	[-]	Proportional gain factor of P-control.
$k_{pz}$	[-]	Stiffness of piezo-actuator.
$k_t$	[-]	Total gain over a series of elements.
$K_i$	[-]	Interfering input gain.
$K_m$	[-]	Modifying input gain.
$\ell$	[m]	Length.
$\ell_g$	[m]	Length of the air gap.
$\ell_i$	[m]	Length of the flux path inside a coil.
$\ell_m$	[m]	Length of the magnet.
$\ell_o$	[m]	Length of the flux path outside a coil.
$\ell_w$	[m]	Wire length of a winding.
$\ell_{w,t}$	[m]	Total wire length of a coil in PM field.
$\ell_{w,a}$	[m]	Active part of the wire in a coil.
$\ell_{w,p}$	[m]	Passive part of the wire in a coil.
$\ell_y$	[m]	Length of the flux path inside a yoke.
$L$	[H]	Self inductance.
$L_a$	[H]	Self inductance of actuator.
$L$	[W/(m <sup>2</sup> sr)]	Radiance of a light source.
$L_{\text{eff}}$	[W/(m <sup>2</sup> sr)]	Effective radiance of a light source.
$m$	[kg]	Mass.
$m_a$	[kg]	Mass of anti-resonator.
$m_m$	[kg]	Mass of mover.
$M$	[-]	Magnification.
$M$	[W/m <sup>2</sup> ]	Radiant emittance.
$M(s)$	[-]	Transfer function of measurement system.
$n$	[-]	Number of, order of TF (integer).
$n$	[-]	Index of refraction.
$N$	[-]	Order number or interferometer constant.
$N_E$	[V <sup>2</sup> /Hz]	Excess, flicker or 1/f noise .
$N_S$	[A <sup>2</sup> /Hz]	Shot noise.
$N_T$	[V <sup>2</sup> /Hz]	Thermal noise.
$\hat{\mathbf{n}}$	[-]	Normal vector on defined surface.
$p$	[-]	Probability.
$P$	[Pa]	Pressure.
$P$	[W]	Power, Intensity.
$P_C$	[W]	Power needed for a voltage over a capacitor.
$P_e$	[W]	Electrical Power.
$P_i$	[W]	Illumination Power of a light source.
$P_{i,c}$	[W]	Total captured illumination power.
$P_{i,t}$	[W]	Total radiated illumination power.

$P_m$	[W]	Mechanical Power.
$P_l$	[W]	Dissipated Power loss.
$P_s$	[-]	Signal power, squared amplitude.
$P_L$	[W]	Power to create a current in an inductor.
$P_v$	[Pa]	Vapour pressure.
$Q$	[-]	Quality factor.
$Q_m$	[-]	Figure of merit.
$q$	[C]	Electric charge.
$q_e$	[C]	Charge of an electron.
$R$	[\Omega]	Resistance.
$R_a$	[\Omega]	Resistance of the actuator.
$R_c$	[\Omega]	Characteristic impedance of a cable.
$R_s$	[\Omega]	Output resistance of the source.
$R_{s,c}$	[\Omega]	Output resistance of the collector source.
$R_{s,e}$	[\Omega]	Output resistance of the emitter source.
$R_{x,y,z}$	[rad]	Rotation around $x, y, z$ axis.
$\mathfrak{R}$	[A/Wb]	Magnetic reluctance.
$\mathfrak{R}_g$	[A/Wb]	Magnetic reluctance of the air gap.
$\mathfrak{R}_i$	[A/Wb]	Magnetic reluctance inside a coil.
$\mathfrak{R}_m$	[A/Wb]	Magnetic reluctance of the permanent magnet.
$\mathfrak{R}_o$	[A/Wb]	Magnetic reluctance outside a coil.
$\mathfrak{R}_t$	[A/Wb]	Total magnetic reluctance of the flux path.
$\mathfrak{R}_y$	[A/Wb]	Magnetic reluctance of the ferromagnetic yoke.
$s$	[-]	Laplace variable.
$S, \mathbf{S}$	[-]	Mechanical strain.
$t$	[s]	Time.
$T$	[°C]	Temperature.
$T, \mathbf{T}$	[N/m <sup>2</sup> ]	Mechanical stress.
$T_c$	[°C]	Curie temperature.
$T$	[s]	Period.
$v$	[m/s]	Velocity.
$v_p$	[m/s]	Wave propagation velocity.
$v_m$	[m/s]	Motion velocity.
$V$	[m <sup>3</sup> ]	Volume.
$V$	[V]	Potential difference, voltage.
$\hat{V}$	[V]	Amplitude of alternating voltage.
$V_a$	[V]	Voltage over an actuator.
$V_b$	[V]	Voltage at the bridge between two switches.
$V_c$	[V]	Voltage of carrier frequency.
$V_{b-e}$	[V]	Transistor base-emitter voltage.

$V_{c-e}$	[V]	Transistor collector-emitter voltage.
$V_{cs}$	[V]	Current sensing voltage.
$V_{c,a}$	[m <sup>3</sup> ]	Active volume of a coil in a magnetic field.
$V_d$	[V]	Voltage over a diode.
$V_g$	[V]	Voltage of a geophone.
$V_C$	[V]	Voltage by charge displacement in capacitor.
$V_L$	[V]	Voltage by current change in inductor.
$V_i$	[V]	Input voltage of an electronic circuit.
$V_m$	[V]	Voltage by the movement of an actuator.
$V_o$	[V]	Output voltage of an electronic circuit.
$V_p$	[V]	Power supply voltage.
$V_R$	[V]	Induced voltage by a current in a resistor.
$V_s$	[V]	Voltage from the source.
$V_t$	[V]	Total series voltage in an electronic circuit.
$V_{th}$	[V]	Base-emitter threshold voltage of transistor.
$V_{Th}$	[V]	Thevenin equivalent voltage.
$x$	[m]	Position, distance.
$x_d$	[m]	Displacement.
$\hat{x}$	[m]	Amplitude of alternating position.
$Z$	[-]	Impedance.
$\beta$	[-]	Current amplification ratio of transistor.
$\gamma$	[-]	Effective fill factor of coil windings.
$\epsilon$	[-]	Control error.
$\epsilon_0$	[As/Vm]	Electric Permittivity (EP) in vacuum.
$\epsilon_r$	[-]	Relative EP, dielectric constant.
$\zeta$	[-]	Damping ratio.
$\vartheta$	[rad or °]	(Capture) angle.
$\vartheta_i$	[rad or °]	Angle of incidence.
$\vartheta_r$	[rad or °]	Angle of reflection.
$\vartheta_t$	[rad or °]	Angle of refraction.
$\lambda$	[-]	Loss factor for magnetic stray flux.
$\lambda$	[m]	Wavelength.
$\lambda_B$	[m]	Bragg grating wavelength.
$\mu_0$	[Vs/Am]	Magnetic permeability of vacuum.
$\mu_r$	[-]	Relative magnetic permeability.
$\varphi$	[rad or °]	Phase angle.
$\varphi_m$	[rad or °]	Phase angle due to movement.
$\Phi$	[Wb]	Magnetic flux.
$\Phi_e$	[C]	Electric charge flux.
$\Phi_m$	[Wb]	Magnetic flux of magnet.

$\Phi_w$	[Wb]	Magnetic flux inside coil winding.
$\Phi_{w,t}$	[Wb]	Total magnetic flux over all windings.
$\Phi_y$	[Wb]	Magnetic flux inside a yoke.
$\rho_r$	[ $\Omega\text{m}$ ]	Resistivity.
$\rho_q$	[C/m <sup>3</sup> ]	Electric charge density.
$\sigma$	[1/m]	Wave number.
$\sigma_x$	[nb]	Standard deviation of variable $x$ .
$\tau$	[s]	Time constant, $RC$ time.
$\tau_d$	[s]	Differentiating time constant (D-control).
$\tau_e$	[s]	Electrical time constant of actuator.
$\tau_i$	[s]	Integrating time constant (I-control).
$\tau_t$	[s]	Taming time constant (D-control).
$\omega$	[rad/s]	Angular frequency.
$\omega_0$	[rad/s]	Angular natural or corner frequency.
$\omega_c$	[rad/s]	Angular unity-gain cross-over frequency.
$\omega_d$	[rad/s]	Differentiating corner frequency of D-control.
$\omega_i$	[rad/s]	Integrating corner frequency of I-control.
$\omega_t$	[rad/s]	Taming corner frequency of D-control.
$\Omega$	[sr]	Solid angle.
$\Omega$	[V/A]	Resistance.

## Abbreviations

ADC	Analogue-to-Digital Converter
AFM	Atomic Force Microscopy
AO	Adaptive Optics
BIPM	Bureau International des Poids et Mesures
BWO	Beam Walk Off
CAD	Computer Aided Design
CAS	Cumulative Amplitude Spectrum
CD	Critical Dimension
CCD	Charge Coupled Device (Camera sensor)
CMMR	Common Mode Rejection Ratio
CMP	Chemical Mechanical Polishing
CPF	Cumulative Probability Function
CPS	Cumulative Power Spectrum
DAC	Digital-to-Analogue Converter
DEB	Dynamic Error Budgeting
DFT	Discrete Fourier Transform
DOF	Depth Of Focus
EMC	Electro-Magnetic Compatibility
EP	Electric permittivity
ESD	Electrostatic Discharge
ESO	European Southern Observatory
FBG	Fibre Bragg Grating
FEM	Finite Element Method
FET	Field Effect Transistor
FFT	Fast Fourier Transform
FRF	Frequency Response Function
FWHM	Full Width Half Maximum
GB	Gain Balancing
GS	Gain Scheduling
GUM	Guide to the expression of Uncertainty in Measurement
HF	High Frequency
HP	High Pass (filter)
IC	Integrated Circuit
IGBT	Insulated Gate Bipolar Transistor
JCGM	Joint Committee for Guides in Metrology
LF	Low Frequency
LP	Low Pass (filter)

LSB	Least Significant Bit
LVDT	Linear Variable Differential Transformer
MA	Moving Average
MEMS	Micro Electro Mechanical System
MFD	Magnetic Flux Density
MIMO	Multiple Input Multiple Output (control)
MOS	Metal Oxide Semiconductor (FET)
MSB	Most Significant Bit
MSD	Moving Standard Deviation
NA	Numerical Aperture
OPD	Optical Path Difference
OPL	Optical Path Length
PALM	Piezoelectric Active Lens Mount
PCB	Printed Circuit Board
PDA	Personal Digital Assistant
PDF	Probability Density Function
PLL	Phase Locked Loop
PLZT	Lead Lanthanum Zirconate Titanate (Piezo)
PM	Permanent Magnet
PSD	Power Spectral Density
PSD	Position Sensitive Detector
PSF	Point Spread Function
PSRR	Power Supply Rejection Ratio
PVDF	Polyvinylidene Fluoride (Piezo)
PTFE	Polytetrafluoroethylene, Teflon
PZT	Lead Zirconate Titanate (Piezo)
PWM	Pulse Width Modulation
RMS	Root Mean Square
SCS	Surface Cross Section
SISO	Single Input Single Output (control)
SMPA	Switched Mode Power Amplifier
SNR	Signal to Noise Ratio
SPM	Scanning Probe Microscopy
STM	Scanning Tunnelling Microscopy
STP	Shielded Twisted Pair
TDM	Time Division Multiplexing
TF	Transfer Function
UTP	Unshielded Twisted Pair
VCO	Voltage Controlled Oscillator
VGM	Vector Gain Margin

WDM

Wavelength Division Multiplexing

# Index

- RC-time, 336  
 “Hooke – Newton” law, 87  
 0 dB level, 83  
 Abbe error, 685  
 absolute standards, 540  
 AC, 49  
 acceleration, 95  
 achromatic doublet, 488  
 adaptive feedforward control, 155, 208  
 adaptive optics, 525  
 aging, 303  
 air foot, 695  
 air mounts, 695  
 air showers, 678, 724  
 air-gap, 235  
 amplitude grating, 506  
 amplitude modulation, 575  
 analogue-to-digital converters, 587  
 angular frequency, 49  
 annular illumination, 521  
 anti-resonance, 121  
 anti-resonator, 703  
 anti-windup control, 212  
 aperiodic, 110  
 aperture plane, 485  
 aperture stop, 492  
 asphericity, 482  
 assist features, 689  
 Astigmatism, 481  
 asymptotically stable, 175  
 at equal amplitude, 76  
 attenuation-band, 335  
 autotransformer, 291  
 azimuthal order, 529  
 balance mass, 695  
 balanced bridge, 560  
 bandwidth, 91, 93, 114, 171  
 barrel, 486  
 base frame, 695  
 batteries, 44  
 BBN criteria, 552  
 beam splitter, 533  
 beam walkoff, 682  
 bias, 402  
 biasing, 402  
 bimorph, 630  
 birefringence, 499  
 bistable, 579  
 bits, 580  
 bitstream, 594  
 black body, 452  
 black-box system identification, 205  
 Bode Sensitivity Integral, 181  
 Bode-plot, 79  
 body, 95  
 branches, 560  
 Brewster’s angle, 499  
 bridge, 428  
 bridge-rectifier, 356  
 brightness, 461  
 Butterworth, 394  
 carrier, 575  
 catadioptric, 473  
 catoptric, 473, 688  
 cats-eye, 660  
 cavity, 504  
 centre-frequency, 663  
 channel, 360  
 characteristic impedance, 325, 597  
 characteristic polynomial, 199  
 charge carriers, 353  
 charge control, 311  
 charge-pumping, 432, 434

- Chebyshev, 394  
chief ray, 493  
chip, 351  
chromatic aberration, 485  
chromatic dispersion, 487  
circular polarised, 500  
clipping, 209, 363  
closed-loop, 146  
closed-loop feedback accelerometer, 629  
coaxial cable, 567  
coercive force, 233  
cogging, 711  
coherence, 456  
coherence length, 459  
coherence time, 458  
collector follower, 365  
Coma, 481  
common collector, 362  
common emitter, 365  
common-mode rejection ratio, 370, 404  
common-mode signal, 367  
Compact Disc player, 6  
complementary sensitivity function, 169  
completer and finisher, 26  
compliance, 86, 95  
constructive interference, 502  
controllability, 131, 205  
converse piezoelectric effect, 296  
corkscrew, 244  
corner-frequency, 335  
cosine error, 676  
creep, 301  
critical angle, 469  
Critical Dimension, 521, 687  
cross-coupling, 195  
cross-over distortion, 364  
Crown glass, 488  
cube-corner retro-reflector, 660  
Cumulative Amplitude Spectrum, 551  
Cumulative Power Spectrum, 548  
Cumulative Probability Function, 544  
Curie temperature, 296  
current source, 44  
cut-off frequency, 335  
cylinder, 483  
damper, 95  
damper-line, 97  
Darlington pairs, 411  
DC, 49  
de-polarisation, 311  
deci-Bel, 81  
decoupling, 120  
Deflectometer, 603  
Delta-Sigma converter, 588  
demagnetisation graph, 233  
departure, 482  
depletion layer, 354  
depletion-mode, 360  
depth of field, 522  
depth of focus, 522, 720  
destructive interference, 502  
diamagnetic, 228  
dichroic coating, 505  
die, 688  
dielectric constant, 329  
differential-mode, 368  
diffraction, 450  
diffraction limited, 518  
digital-to-analogue converter, 590  
dioptric, 473  
dioptric, 473  
direct band-gap, 455  
direct piezoelectric effect, 296  
discrete components, 350, 409  
discrete Fourier transform, 71  
displacement, 95  
distortion, 485  
divergence, 219  
Doppler shift, 667  
Dose-control, 522  
double logarithmic, 80  
double-telecentricity, 493  
Drain, 359

- drift, 402  
dual-ended, 435  
dual-slope ADC, 587  
Dutch school of mechatronics, 20  
duty-cycle, 424  
Dynamic Error Budgeting, 52, 544  
eddy-current, 258  
eigendynamics, 109  
eigenfrequencies, 121  
eigenmodes, 118, 125  
eigenvalues, 109, 125, 199  
electric load, 324  
electric motor, 217  
electric permittivity, 329  
electric permittivity in vacuum, 39  
electric signals, 48  
Electro Magnetic Compatibility, 409  
electrodes, 43  
electromagnetic actuator, 217  
Electromotive Force, 44  
electronically commutated, 251  
electrostatic force, 39  
elliptical polarised, 500  
emitter follower, 362  
enhancement-mode, 360  
entry-pupil, 492  
etendue, 464  
exit-pupil, 492  
f-number, 519  
Fabry-Perot interferometer, 504  
fading, 693  
far field, 507  
Faraday shield, 569  
Fast Fourier Transform, 71  
features, 687  
feedback, 139  
feedforward, 139  
Fibre Bragg Grating, 617  
fiducials, 718  
field curvature, 485  
field lines, 39  
field plane, 485  
figure of merit, 254  
finesse, 505  
first-order, 108  
flare, 481  
flexure-scanner, 309  
flicker noise, 554  
Flint glass, 488  
floating, 382, 432  
floor vibrations, 548  
focal plane, 485, 710  
force frame, 696  
four quadrant, 417  
four-quadrant detector, 601  
Fourier transform, 65  
fourth-order, 125  
free electrons, 353  
frequency combs, 671  
frequency domain, 52, 72  
frequency response function, 101  
frequency spectrum, 52  
frequency-to-digital conversion, 579  
fringes, 659  
Full-width at half-maximum, 457  
Gain Balancing, 727  
gain margin, 171  
Gain Scheduling, 727  
gain-bandwidth product, 401  
gang of four, 169  
gang of six, 168  
Gate, 359  
Gaussian distribution, 545  
geometric optics, 450  
geophone, 623  
Gray code, 582  
grey-box system identification, 205  
grid plate, 726  
ground loop, 569  
group delay, 395  
H-bridge, 435  
half-pitch, 688  
Hall effect, 444  
hard-switching, 438  
Hartmann sensor, 533  
Heisenberg uncertainty principle, 541

- heterodyne interferometry, 662  
holes, 353  
homo-polar, 277  
homodyne detection, 578  
homodyne interferometer, 657  
hot electrons, 358  
hydraulic linear motors, 13  
Hysteresis, 302  
hysteresis operators, 302  
  
I-control, 165, 174  
idle current, 362  
illumination power, 460  
image-space telecentric, 495  
impedance, 54  
in Abbe, 686  
in parallel, 272  
in series, 274  
independent, 126  
indirect band-gap, 455  
inductor, 260  
inertial velocity sensor, 623  
innovator and creator, 26  
input-shaping, 152  
instrumentation amplifier, 573  
Insulated Gate Bipolar Transistor, 422  
Integrated Circuits, 9  
integrator windup, 212  
intensity, 59  
interference, 450, 501  
interfering input, 542  
interferometer, 503  
interferometer constant, 658  
intermediate image, 490  
inverse piezoelectric effect, 296  
iris diaphragm, 492  
irradiance, 60  
iterative learning control, 155  
  
jerk, 279  
  
k-one factor, 521  
Kalman-filter, 205  
Kalman-gain, 207  
  
Laplace domain, 73  
Laplace plane, 106  
laser, 456  
latch, 582  
LCR-filter, 340  
lead-lag compensation, 163  
lead-network, 162  
Least Significant Bit, 581  
lensmakers equation, 478  
level-shifter, 411  
levelling, 720  
limit-cycling, 212, 411  
linear polarised, 499  
Lissajous plot, 646  
long tailed pair, 369  
long-lead items, 32  
long-stroke, 19, 711  
loop shaping design, 170  
Lorentz force, 222  
low-pass filter, 189  
  
machining, 19  
magnetic bearing, 141  
magnetic interference, 566  
magnetomotive force, 226  
majority carriers, 358  
mass dilemma, 715  
measurand, 539  
measurement accuracy, 540  
measurement error, 539  
measurement precision, 540  
measurement resolution, 541  
measurement uncertainty, 540  
Mechanical amplification, 308  
meridional plane, 484  
metrology, 537  
metrology frame, 684, 695  
metrology loop, 685, 716  
Miller capacitor, 397  
minimal realisation, 205  
minority carriers, 358  
mirage, 526  
modal analysis, 129  
mode-shapes, 125

- model-based, 149  
modified Edlén, 676  
modifying input, 542  
modulation, 575  
Moore's law, 691  
Most Significant Bit, 581  
mover, 265  
multi-lens array, 533  
multi-sines, 78  
Multiple Input Multiple Output, 195  
N-material, 353  
negative stiffness, 140, 270  
noise, 49  
normal dispersion, 487  
normal distribution, 545  
notch filter, 151  
NTC, 558  
Numerical Aperture, 519  
Nyquist frequency, 586  
Nyquist plot, 83  
object-space telecentric, 495  
observability, 131, 205  
observer, 205  
off-the-shelf, 33  
offset, 401  
one-over-f ( $1/f$ ) noise, 554  
open-loop, 81, 140, 159  
open-loop control, 143  
Operational Amplifiers, 361  
optical axis, 474  
optical flats, 675  
Optical path length, 467  
optical pick-up unit, 5  
optical throughput, 464  
optimal control, 192  
order and delivery, 33  
over-constrained, 88  
over-hung, 250  
overactuation, 730  
overlay, 691  
P-control, 159  
parallel-resonant, 341  
paramagnetic, 228  
parasitic, 327  
paraxial rays, 475  
pass-band, 335  
PD-control, 156, 161  
pedestal, 695  
pentafoil, 531  
periodic error, 673  
Perovskite structure, 296  
phase, 50  
phase grating, 508  
phase margin, 171  
phase selective detection, 578  
photons, 451  
physical optics, 450  
PID-control, 156, 165  
piezo-gain, 300  
Piezo-stack actuators, 306  
Piezoelectric Active Lens Mount, 702  
piezoelectric scanner, 309  
Piezoelectric tube actuators, 305  
piezoresistivity, 619  
Pincushion, 486  
pink noise, 52  
plant, 138  
platforms, 30  
point source, 461  
point-symmetric, 661  
polarisability, 329  
polarisation, 57, 296, 497  
polarisation mixing, 673  
polariser, 499  
polarising beam splitter, 664  
pole pieces, 235  
pole-splitting, 400  
pole-zero cancellation, 144  
poles, 106, 139  
poles and zeros, 73  
potential difference, 41  
potentiometer, 323  
Power Spectral Density, 548  
power supply rejection ratio, 404  
Pre-loading, 307  
Precision Engineering, 19

- Preisach model, 302  
primary windings, 292  
printed circuit board, 328  
prism, 530  
Probability Density Function, 544  
product creation process, 26  
product manager, 27  
propagation, 57  
proportional, 158  
proportional control, 159  
Proportional Differential control, 161  
proximity detector, 599  
pulse-width modulation, 425  
pupil, 491  
pupil plane, 485, 494  
purple fringing, 487  
Push Pull class AB, 364  
Push-Pull class B, 364  
quad-cell, 601  
quadrafoil, 531  
quadrature detection, 669  
quality factor, 106  
quantisation error, 581  
quarter wave plate, 663  
R-2R ladder-network, 590  
race track coil, 248  
radial order, 529  
radiance, 456, 461  
radiant emittance, 461  
radiant flux, 460  
radiant intensity, 60  
random errors, 540  
ray tracing, 479  
RC-filter, 335  
reactive, 56, 330  
reactive impedance, 328  
real value, 539  
recombination, 354  
reflection, 61, 449  
refraction, 449  
Reluctance, 226  
reluctance force, 257  
remnant flux density, 233  
requirement budgeting, 28  
Resistivity, 38  
resonance, 64  
resonant-mode, 438  
reticle, 10, 522  
reticle stage, 689  
reverse-recovery time, 439  
right hand, 244  
road mapping, 30  
robustness, 147  
roll-off, 99  
roll-off frequency, 93, 165  
root of the sum of squares, 547  
rotation, 219  
safety ground, 570  
sagittal plane, 484  
sagittal rays, 484  
Sallen-Key, 391  
sample-and-hold, 584  
samples, 583  
saturation, 209  
scanner bow, 309  
Schmitt trigger, 372, 425  
Schottky diodes, 439  
second-order, 102  
secondary breakdown, 422  
secondary trefoil, 531  
secondary windings, 292  
seismic mass, 624  
selective amplifier, 576  
self-inductance, 260  
sensitivity function, 169  
serial scanner, 309  
series-resonant, 344  
servo-system, 6, 156  
servo-valves, 13  
settle time, 728  
Shack-Hartmann wavefront sensor, 533  
shadow mask, 12  
Shielded Twisted Pair, 570  
short-stroke, 711  
side-bands, 577

- Silicon Repeater, 2, 9  
 simulated, 166  
 Single Input Single Output, 143  
 single-ended, 428  
 single-ended Class A, 363  
 single-phase, 437  
 skyhook, 699  
 slew-rate, 209, 401  
 slit, 689  
 smart-disk, 702  
 snap, 282  
 solenoid, 333  
 solid-state lamps, 455  
 Source, 359  
 space, 195  
 spatial coherence, 458  
 spatial frequency, 516  
 specific peak force, 714  
 speckle, 457  
 Spectrum Analyser, 71  
 Spherical Aberration, 481  
 split-frequency, 663  
 spring, 95  
 spring-line, 96  
 stability, 140  
 stable, 140  
 standard deviation, 545  
 star configuration, 446  
 state feedback, 202  
 state-space, 137, 156, 195  
 state-variables, 195  
 stator, 265  
 stimulus, 74  
 stops, 519  
 strain gage, 558  
 stray flux, 237  
 Strehl ratio, 519  
 successive-approximation, 588  
 super-capacitors, 332  
 super-spring, 201  
 Surface Mount Devices, 328  
 switched-mode power amplifiers, 418  
 switched-mode power supply, 295  
 synchronous demodulation, 576  
 System Identification, 204  
 systematic effects, 52  
 systematic errors, 540  
 systems engineer, 27  
 tacho-generator, 213  
 tamed PD-control, 163  
 tangential plane, 483  
 tangential rays, 484  
 temporal coherence, 458  
 temporal frequency, 49  
 temporal or spatial variability, 48  
 terminals, 43  
 thermal-centre, 608  
 thin-lens equation, 478  
 three-phase, 251  
 throughput, 692  
 Time Division Multiplexing, 622  
 time domain, 52, 72  
 total quality, 32  
 trace, 479  
 traceability, 537, 717  
 traceable, 540  
 Traité de la lumière, 505  
 transconductance, 381, 413  
 transfer function, 101  
 transformer core, 291  
 transimpedance, 380  
 transmissibility, 86, 95  
 transmission-line, 596  
 trefoil, 531  
 triangulation, 603  
 tripod-scanner, 309  
 tuned-mass damper, 703  
 two degree of freedom control, 214  
 ultimately, 112  
 unambiguously measurable, 24  
 unbound, 218  
 under-hung, 251  
 unity-gain cross-over frequency, 83, 158  
 Unshielded Twisted Pair, 566  
 V-model of systems engineering, 23

- vector margin, 171  
velocity, 95  
Vernier, 648, 719  
vias, 691  
vibration isolation, 115  
vibration isolation system, 548, 695  
Video Long Play Disk, 2  
vignetting, 496  
virgin curve, 232  
virtual, 477  
virtual ground, 377  
voltage controlled, 311  
voltage source, 43  
wafer chuck, 689  
wafer fab, 695  
wafer stage, 13, 689  
wafer table, 689, 726  
waterbed effect, 180  
wave-guide, 596  
wavelength, 57  
Wavelength Division Multiplexing, 621  
wavelength tracker, 677  
Weiss domains, 231, 297  
Wheatstone bridge, 559  
white noise, 52, 78, 553  
Zernike polynomials, 528  
zero-stiffness, 19, 705  
zeros, 139  
zoom lens, 490

This page intentionally left blank