IntechOpen

ICT - Energy
Concepts Towards Zero - Power Information and Communication Technology

Edited by Giorgos Fagas, Luca Gammaitoni,
Douglas Paul and Gabriel Abadal Berini

ICT - ENERGY - CONCEPTS TOWARDS ZERO - POWER INFORMATION AND COMMUNICATION TECHNOLOGY

Edited by **Giorgos Fagas, Luca Gammaitoni, Douglas Paul and Gabriel Abadal Berini**

ICT - Energy - Concepts Towards Zero - Power Information and Communication Technology
http://dx.doi.org/10.5772/55410
Edited by Giorgos Fagas, Luca Gammaitoni, Douglas Paul and Gabriel Abadal Berini

Contributors

James Rohan, Gabriel Abadal, Javier Alda, Jordi Agustí, Igor Neri, Antonio Moschitta, Helios Vocca, Francesco Cottone, Francesco Orfei, Giulia Orecchini, Naser Khosro Pour, Douglas J. Paul, Luca Gammaitoni, Giorgos Fagas

Notice

Statements and opinions expressed in the chapters are these of the individual contributors and not necessarily those of the editors or publisher. No responsibility is accepted for the accuracy of information contained in the published chapters. The publisher assumes no responsibility for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained in the book.

First published in Croatia, 2014 by INTECH d.o.o.
eBook (PDF) Published by IN TECH d.o.o.
Place and year of publication of eBook (PDF): Rijeka, 2019.
IntechOpen is the global imprint of IN TECH d.o.o.
Printed in Croatia

Legal deposit, Croatia: National and University Library in Zagreb

Additional hard and PDF copies can be obtained from orders@intechopen.com

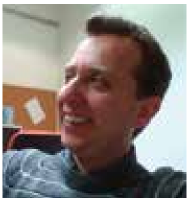
ICT - Energy - Concepts Towards Zero - Power Information and Communication Technology
Edited by Giorgos Fagas, Luca Gammaitoni, Douglas Paul and Gabriel Abadal Berini

p. cm.

ISBN 978-953-51-1218-1

eBook (PDF) ISBN 978-953-51-6358-9

Meet the editors



Dr. Giorgos Fagas received his PhD degree in Physics from Lancaster University (UK) in 2000 and completed his MBA degree at University College Cork in 2012. He is currently the EU Programme Coordinator at Tyndall National Institute, Ireland. As senior researcher at Tyndall, he has coordinated activities in nanoelectronics and energy-efficient electronics. From 2000 to 2003, Giorgos was research fellow at the Max-Planck-Institut-PKS in Dresden and Humboldt Fellow at the University of Regensburg, Germany. His work has been published in over 50 peer-reviewed articles spanning: atomic scale sub-10 nm transistor designs and electronic device evaluation, implementations of first-principles electronic structure methods to electron transport, device physics and fundamental mesoscopic phenomena. Giorgos has also edited a reference book on molecular electronics.

Dr. Luca Gammaitoni, is Professor of Experimental Physics at the University of Perugia, in Italy and the director of the Noise in Physical Systems (NiPS) Laboratory. He is also the founder of Wisepower srl a university spin-off company. He obtained the PhD in Physics from the University of Pisa in 1990. Since then he has developed a wide international experience with collaborations both in Europe, Japan and the USA. His scientific interests span from noise phenomena in physical systems to non equilibrium thermodynamics and energy transformations at micro and nanoscale. He authored over 200 papers on top-level scientific journals and few books. He is also the author of 10 patents. His papers have been cited more than 10.000 times. H factor=43.

Dr. Douglas Paul has degrees in physics from the University of Cambridge and spent 15 years as a researcher there before moving to the University of Glasgow, U.K.. He is Professor of Semiconductor Devices and Director of the James Watt Nanofabrication Centre at Glasgow and presently sits on a number of U.K. Government Scientific Advisory Committees. His research includes nanoelectronics, quantum devices, Si photonics and energy harvesting including thermoelectrics. He has published over 250 articles in refereed journals and conference proceedings.

Dr. Gabriel Abadal received his degree in physics in 1991 and his PhD in electrical engineering in 1997 from the UAB. Since 2002 he is an Associate Professor in the Electronics Engineering Department of the UAB, where he currently leads the NANERG LAB group (http://grupsderecerca.uab.cat/nanerglab/). During the last 20 years he has participated in more than 30 national and European research projects, he has contributed to more than 160 conferences and he has published more than 75 papers in journals (H-index: 15). Since 2006, he focused his activity in the application of MOEMS and NOEMS to harvest energy from ambient vibrations and electromagnetic radiation sources at the nanoscale. His actual interest is oriented towards the study of graphene and other 2D materials as basic components of the future nanoenergy harvesting strategies.

Contents

Preface

This book deals with the relationship between the fast growing field of Information and Communication Technology (ICT) and the energy necessary to power this growth. In the last twenty years there has been an explosion of popularity in portable ICT devices.

The electronic technology, fostered by a fast-developing semiconductor industry, has made impressive progress in reducing the size of the basic element of the machinery of computation: the transistor. The CMOS field effect transistor (FET) is just the recent acquisition in this chain of products, of smaller and smaller size. Such a continuous decrease has produced a rapid increase in the computational density of standard microprocessors. Such a continuous growth from the scaling to smaller dimensions, however, is now facing an end due to the relevant heat production as a side effect of the computation process.

Further development of the technology in portable devices is presently at stake if we do not find a way to bridge the gap between the amount of energy required to operate such devices and the amount of energy available from portable/mobile sources. The only viable solution appears to be to attack the gap from both sides, i.e. to reduce the amount of energy dissipated during computation and to improve the efficiency in energy harvesting technologies.

In this book we list a number of contributions that address the various aspects of these two approaches, starting from the description of the fundamental limits in energy dissipation attainable in the basic elements of computing devices and comparing them to the actual figures from the present CMOS technology. Energy harvesting is discussed in great detail. We present contributions from a team of scientists that are actively involved in the challenge to make energy harvesting more efficient and reliable and represent the forefront of the scientific and technological research in the field of micro and nanoscale energy harvesting devices.

This book is realized with the contribution of the project "ZEROPOWER: Co-ordinating Research Efforts Towards Zero-Power ICT" funded under the Future and Emerging Technologies (FET) programme of the Seventh Framework Programme for Research of the European Commission (Grant Agreement n. 270005).

Giorgos Fagas
EU Programme Coordinator
Tyndall National Institute - University College Cork
Lee Maltings, Dyke Parade, Cork, Ireland

Luca Gammaitoni
NiPS Laboratory, Università di Perugia, Italy

Douglas Paul
University of Glasgow, School of Engineering, Rankine Building, U.K

Gabriel Abadal Berini
Departament d'Enginyeria Electrònica, Escola d'Enginyeria
Universitat Autònoma de Barcelona , Barcelona, Spain

Chapter 1

A Research Agenda Towards Zero-Power ICT

Gabriel Abadal Berini, Giorgos Fagas,
Luca Gammaitoni and Douglas Paul

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57347

1. Introduction

Societies reliance and use of Information Communications Technology (ICT) is increasing. It is estimated that 2% of all energy consumption is now the result of ICT use. The energy consumption and carbon dioxide emission from the expanding ICT use, however, is unsustainable and will impact heavily on future climate change. Methods are required to make ICT technology more energy efficient but also the development of new self-powered, energy-harvesting technologies that would enable micro- and nano-scale systems that consume Zero-Power through the harvesting of waste energy from the environment are also required. Autonomous sensors for temperature and pollution monitoring are also key for SMART metering to reduce energy consumption in domestic and industrial environments. Zero-Power autonomous sensors for healthcare applications have the potential to change the expensive reactive healthcare market to a cheaper and more effective point-of-care diagnostic system. Such healthcare sensors also have the potential to radically change the care of the elderly to a more sustainable and scalable automated monitoring rather than present expensive labour intensive methods.

In this chapter, a brief introduction is given on the challenges and possible solutions followed by an analysis on the technological and market impact as well as ethical and societal issues arising from the use of Zero-Power ICT solutions for reducing the energy required in new services and products.

2. Background

2.1. Energy efficient ICT

ICT has become a strategic sector in the world economy. Its impact on cultural and social development is already paramount and it will keep growing in the foreseeable future. State of the art ICT is presently based on digital devices whose functioning is currently dominated by power dissipated which produces heat. This is a major problem for a number of reasons:

1. *Economic and social reasons*. Energy efficient ICT, that is, operating ICT devices with reduced amount of energy, is presently considered an objective of extremely high economic relevance. According to the SMART 2020 [1] study, "the share of ICT on the worldwide energy consumption today is in the range of 2-5%. Given that the use of ICT will further increase and the overall energy consumption will hopefully decrease due to the help of ICT and other measures, it is expected that the share of ICT on the worldwide energy consumption will grow in the future. Carbon dioxide emissions from the use of ICT are therefore presently increasing. Hence, it becomes more and more important to consider and improve the energy efficiency of ICT. In the short term, it will be an obvious and practical solution to better exploit the potential of technologies that already exist or are currently in the making. On the long term, new and disruptive ideas will be needed" [2].
2. *Technological reasons*. In the last forty years the semiconductor industry has been driven by its ability to scale down the size of the CMOS Field Effect Transistor switches, the building blocks of present computing devices, and to increase computing capability density up to a point where the power dissipated in heat during computation has become a serious limitation. In fact, the power density in modern day chips is several tens of W/cm^2 which implies a surface hotter than high-power hot plates. According to the International Technology Roadmap of Semiconductors [3] the limits imposed by the physics of switch operation will be the roadblock for future scaling in the next 10-15 years. The limit on the minimum energy per switching is set at approximately $3\ k_BT \ln(2)$ (approx 10^{-20} J at room temperature) [4].

Power dissipated versus switching speed of devices have been characterized since the 1970s [5] by a linear scaling rule where micro-fabrication capabilities, through the replacement of bipolar transistors with CMOS, allowed the continuation of the exponential increase trend in information processing capability. This has been known as Moore's law. However, since 2004 the Nanoelectronics Research Initiative[1], a US based consortium of Semiconductor Industry Association companies, has launched a grand challenge to address the fundamental limits of the physics of switches. Such limits are mainly represented by the minimum energy and minimum time, required to operate a switch. With the present estimate of the minimum energy required in current CMOS technology (with the Field Effect Transistor channel scaled down

[1] The Nanoelectronics Research Initiative (http://www.src.org/program/nri/) was formed in 2004 as a consortium of Semiconductor Industry Association (SIA) (www.siaonline.org) companies to manage a university-based research program as part of the Semiconductor Research Corporation (SRC) (www.src.org).

to 1.5 nm, switching speed of about 40 fs) the resulting power density for these switches at maximum packing density would be on the order of 1 MW/cm^2. This is comparable to the power density in a rocket nozzle and is orders of magnitude larger than what is presently technologically manageable. Thus the amount of energy dissipated through heat is presently the major roadblock for continuing the increase in computing performance.

3. *Scientific reasons*. Presently the main effort to overcome the technological limitations is aimed at cooling the chips by removing the heat produced during computation. Specific attention goes to the charge transport on one hand and on the other hand on reducing the voltage operating levels up to the point of not compromising the error rate due to voltage fluctuations. Such a strategy has produced some interesting results[2] however it is clearly coming to an end due to the unsustainable energy input requirement. There are attempts to look at the problem from a more fundamental point of view by addressing the basic mechanisms behind the heat production and the role of fluctuations arising by lowering the threshold voltages.

2.2. Micro- and nano-scale wireless sensors

MEMS (Micro Electro-Mechanical Systems) and NEMS (Nano Electro-Mechanical Systems) technology has made significant progress in the last ten years and a new potential in distributed sensing and actuating devices is now approaching the market. Deployment in Wireless Sensor Networks (WSN) has aroused a lot of interest both in academia and industry. There is an increasing demand for ambient intelligence devices, various kinds of sensor networks for safety and environmental monitoring and for monitoring of the health of humans and animals.

Nowadays, thanks to advances in MEMS technologies, it is feasible to fabricate cheap and small sensor nodes that have not only sensing, but also data processing and communicating capabilities. WSNs differ from traditional wireless networks by features like the number of nodes, constraints on energy consumption, computation and memory. Obviously these unique characteristics bring new challenges to the research activity. One of the advantages of wireless communication is the easiness in deploying the sensors without the need of wiring and fixed positioning, thus reducing installation and maintenance cost. Furthermore, in some hostile environment it is difficult or even impossible to physically connect the sensors; therefore wireless communication is the only feasible solution. Due to the heterogeneity of potential application for WSN specific objective and constraints have to be taken into account.

These devices all need distributed powering systems. Presently this means wired power-grids, batteries or RF-sources, however all these solutions present some drawbacks. Wiring is expensive, adds weight and is subject to high failure rates in devices subjected to repeated motion. Traditional batteries are not a viable solution to the powering of such devices mainly because they have to be replaced once exhausted and the cost of replacing the batteries is many orders of magnitude greater than the complete cost of the systems. Alternative solutions based on micro fuel cells and micro turbine generators are also not suitable. Both involve the use of

[2] See e.g. the Aquasar computer installed on 2010 at the Swiss Federal Institute of Technology (ETH) Zurich.

chemical energy and require refuelling when their supplies are exhausted. Thus the goal of powering such devices with energy harvested from the ambient has been in recent years the subject of a great research effort.

If one can realize an energy harvester with a capacity to deliver power of 100 μW, it would open up a large number of applications. Also, the availability of this kind of power sources would boost the development of even lower power devices, leading to the vastness of autonomous nanoscale ICT systems for implants and in-vivo health monitoring, environmental warning and hazard preventing networks and for other safety measures. In a wider context, electronic devices currently account for 15% of household electricity consumption, but their share is rising rapidly, mainly due to growing demand in Africa and the developing world. Next to the need for more secure and greener energy supplies at a large scale, immediate action has to be taken to employ alternative energy sources and reduce power consumption in consumer electronics at all levels.

3. Matching problems with solutions

Improving energy efficiency in ICT and powering networks of small wireless sensors are two important fields of active research that sit on a common scientific background: the management of energy from the micro- and nano-scales to the system level in conjunction with the processing of information.

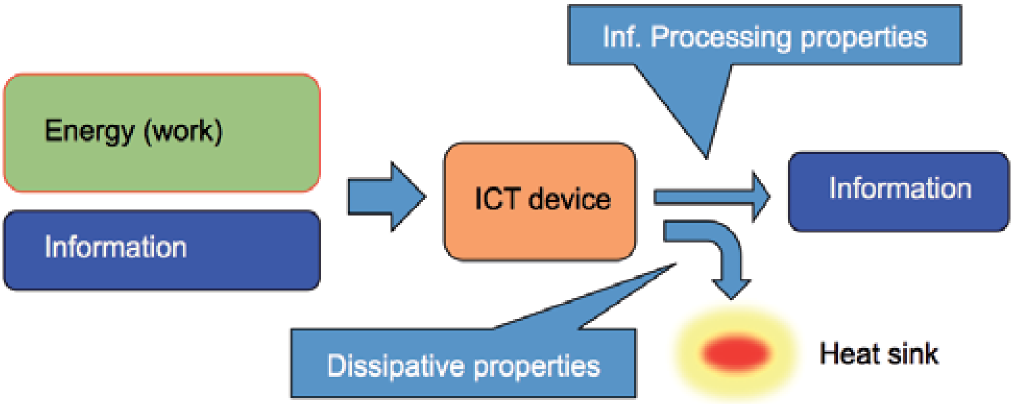


Figure 1. An ICT device is a machine that inputs information and energy (under the form of work), processes both and outputs information and energy (mostly under the form of heat).

Relevant scientific breakthroughs are needed on this topic for making progress in the two fields. Specifically, a new approach is required to the energy management of physical mechanisms at the nanoscale with the aim of setting the bases for a new thermodynamics of ICT devices. In this perspective an ICT device has to be considered as a machine that inputs information and energy (under the form of work), processes information and outputs information and energy (mostly under the form of heat). This is shown schematically in Figure 1. Energy efficiency is usually defined as "the percentage of energy input to a device that is consumed in useful work and not wasted as useless heat". This definition, however, does not apply when we deal with processes at the nanoscale. Moreover the very basic mechanism behind energy dissipation requires a new definition when non-equilibrium processes involving only a few degrees of freedom are considered: the dream of highly efficient devices has to deal with a rethinking of both energy and information dissipation processes.

The long-term aim of research activities addressed here is to make possible low power ICT devices with a significant impact on energy efficiency on a much broader scale. Clearly, a new generation of energy efficient ICT devices has to deal with energy transformation processes at the nanoscale and allow for efficient power management of the realized system. This is undoubtedly a multidisciplinary task where competences from fields as diverse as physics, computer science, electronic engineering and mechanical engineering are brought together in a coordinated effort. This is evident in this remaining chapters of the volume where more technical details are presented. There are also a number of roadmaps towards providing Zero-Power technology for specific applications [6-10].

4. Standards and metrology

To enable comparison of different technologies, a key area that requires development is metrology for energy harvesting systems especially at the micro- and nano-scale. At present there are a lot of bold claims of energy harvesting devices that will save the planet and produce amounts of energy that defy the laws of physics. If consumers and markets are to have the belief and confidence in energy harvesting technologies to allow them to be implemented and used then standards and measurements that allow accurate comparison and benchmarking of the technologies are required.

The issues for metrology are that the input sources of energy come from a large range of sources. As discussed later in the volume the energy to be harvested can be in the form of kinetic, potential, electromagnetic, thermal or chemical energy and for ICT it will be converted into electrical energy. For each application, a different load impedance is likely and the electrical energy needs to be converted and impedance matched to the load. Standards are therefore required for a range of different potential applications to highlight the different areas of energy harvesting.

As an example, the efficiency of thermoelectric devices requires the accurate measurements of electrical conductivity and thermal conductivity. Thermal conductivity is already extremely difficult to measure in bulk systems with uncertainties of around 50% due to the difficulty of heat transport through any object that touches the item to be measured. As thermometers are required to measure the temperature, the act of measurement adds uncertainty to the measurement through the parasitic loss of heat. At the micro- and nano-scales, the measurement is far more complicated and difficult and new techniques and ideas are required. For vibrational energy harvesting the issues will be related to what frequency and bandwidth should be used to compare devices.

There is already a European Metrology Research Programme on energy harvesting technologies (http://www.emrp-metrology-for-energy-harvesting.blogspot.com/). The ZEROPOWER network (http://www.zero-power.eu/) has already discussed a number of collaborative projects on energy harvesting standards and it is clear for future energy harvesting that metrology and standards are at the core of providing robust and quantitative performance data for energy harvesting technologies.

5. Potential impact, applications and markets

The impact of realising low-power electronics in the energy budget is huge. As global energy demand increases, ICT and consumer electronics (CE) account for one of the fastest growing sectors regarding energy consumption. The International Energy Agency (IEA) projects that by 2030 the global energy use by just residential electronic gadgets could rise to 1,700TWh. This is around 40% of the generation capacity of the largest electricity producer in the world (USA). In fact, it would require at least a dozen power plants with 1-2GW output to accommodate this trend which if placed in a single location would make that country the third largest electricity producer. In the developed world, roughly one third of the annual electricity bill from residential ICT and CE is charged to computers, peripherals and other mobile devices; audio-visual equipment accounts for the other two third (estimates based on IEA data and an independent study by the Consumers Electronics Association-USA).

Even greater benefits may derive from new applications. The Smart 2020 study [1] indicates that "ICT's largest influence will be by enabling energy efficiencies in other sectors, an opportunity that could deliver carbon savings five times larger than the total emissions from the entire ICT sector in 2020." Indeed, miniaturised electronic systems applied in ambient intelligence, point-of-care diagnostics, chemical warfare security, logistics and supply-chain control can potentially achieve large cost/energy savings with additional huge societal impact.

The OECD Health Data show that an average of 8.9 % of GDP in developed countries is spent on healthcare costs. Over the last 50 years, healthcare spend has outpaced GDP growth by about 2 percentage points a year in most OECD countries and there are few signs that this trend will slow. Advances in e-Health, the healthcare practices supported by ICT, are a very promising route to reduce the bill. For example, point-of-care devices to diagnose Acute Coronary Syndrome can yield results around 1.0 to 1.5 hrs earlier than analysis in the central laboratory, allowing for earlier intervention or rule out to a step down unit. At a cost of €1,300-€6,500 per hour of emergency department bedtime, the shorter turnaround time provides major savings. In m-Health, mobile electronics and communication technologies are utilised to deliver solutions in prescription drugs monitoring and in remote diagnosis and even treatment for patients who do not have easy access to a physician. Based on ubiquitous intelligence from energy-efficient miniaturised sensors (e.g., in wearable/textile integration), remote health monitoring devices that track and report patients' conditions are possible. Such

ICT solutions are urgently needed taking into account Europe's aging population and the even more demanding constraints on resources and patient empowerment. Current estimates on cost savings from m-Health for chronic diseases in OECD countries are valued at €227 - €273 billion. m-Health is already around a €2 billion market according to CSMG, and it is expected to grow over the next five years at a 25 percent CAGR (compound annual growth rate). A McKinsey&Company report estimates an untapped consumer-led market potential of up to several tens of billions Euros.

Estimates of the current WSN market, which is expected to grow rapidly, are at €1-2 billions. The market of energy harvesting ICT devices alone is estimated close to €1 billion in 2011, to grow to around €5.7 billion by 2021. This is based on "250 million sensors powered by an energy harvester (at an average price of $6 per harvester), and by then numerous consumer electronic devices including laptops, ebooks and cell phones" (IDTechEx Report). However, market fragmentation owing both to enabling technology and end-user application makes it very difficult to reveal the full potential of the market. The healthcare sector was mentioned above. Another short-term windfall impact is expected from applications in energy efficient buildings. For example, just the wireless-enabled HVAC sensors (heating, ventilation and air-conditioning) tapped into the building automation systems will have a market of €150 million and will result in multi-billion euro savings (estimates of 40%-50% energy savings have been reported).

The potential benefits from the research and development of low-power nanoelectronic components and autonomous sensors stretch out to many other research areas. Detailed knowledge of the dynamics of energy/information carriers and the realisation of appropriate channels is of paramount importance for Beyond Moore technologies. In the short-to-medium term, the required advances in energy management concepts, ranging from managing dissipation and fabricating thermodynamically efficient devices at the nanoscale to designing materials and circuits for more efficient electronics, will have diverse impact in metrology, nanofabrication, characterisation and modelling, materials research and smart grid applications.

6. Potential ethical and societal issues

A first general issue, inherent to the Zero-Power ICT concept, which will have a clear ethical and societal impact is related to the contribution of the ICT to climate change. It is nowadays well established that around 2% of the global emissions of carbon dioxide are due to the manufacture, use and disposal of ICT devices and systems. This percentage can increase to 3% in 2020 if the efficiency of ICT is not improved. Zero-Power technologies, including here the binomial **ultralow power electronics – energy harvesting strategies**, will have a double impact in this issue: from one hand they can stop the growth and even reduce the 2% contribution to CO_2 emissions, but on the other hand, they can help to enable the reduction of the remaining 98% of the total emissions produced by non-ICT actors. The main way that ICT will be used in improving energy efficiency is through adding smartness or intelligence to the functionalities/systems in which energy efficiency is required to be improved. Thus, new concepts such as **smart city**, **smart work**, **smart grid, intelligent transport** or smart/intelligent whatever will arise that require autonomous sensing systems that consume no power. Zero-Power technologies will have the opportunity to demonstrate their capabilities to provide smartness/ intelligence to all these functionalities/systems in an energy efficient way.

So, a first type of application will be focused in this so called "ICT for the energy efficiency improvement of ICT and non-ICT systems". In this class of applications we can include:

- Devices and systems for the improvement of energy distribution, consumption and management in general at home, industries and civil buildings and infrastructures. **Smart grid** concept and extensions of the concept to other kinds of energy as gas or renewables (solar, wind).
- Devices for the energy efficiency improvement of ICT systems and equipment: from tiny integrated microsystems to large data centres.

A second issue is related to the improvement the quality of life of the society. Although applications in the previous point have also an incidence on the population quality of life, this impact is expected to be plausible in a medium to long term. Here, a set of applications which have a short term, more direct impact to the quality of life are listed:

- Health related applications such as Body Area Networks (BAN) technologies for health monitoring of elderly, sick, disabled or newborn people.
- Animal tracking and monitoring such as WSN for remote control of position and vital constants of livestock.
- Ambient monitoring using WSN for pollution monitoring in big cities or industrial regions, for fire prevention in forests.
- Geo-atmospheric monitoring employing WSN for monitoring, prediction and prevention of natural catastrophes such as flooding, hurricanes, snow slides, earthquakes or tsunamis.
- Security improvement and accident prevention in cars where systems are already implemented in aircrafts that can minimize the probability of collision as well as detection of individual or collective behavioural patterns that can lead to a dangerous situation.
- Goods tracking with smart active RFID technologies for the improvement of manufacturing and distribution of goods at the industry and distributor level, but also for helping the user in quotidian buying and consuming activities.
- Circulation improvement of public and private transport in cities using smart navigation systems for dynamic calculation of efficient itineraries and smart parking search systems.

In general, energy harvesting and ultralow power electronics will act as enabling technologies in all previous applications. Thus, for instance, most of the previous applications are strongly connected to efficient wireless communication technologies and other related technologies such as **wireless sensor networks** (WSN). So, it is expected that WSN technology will not be deployed in real applications until both energy harvesting and low power electronics will provide good solutions for self-powering WSN nodes and for decreasing energy consumption of the sensing and communication functions respectively.

However, it is important to educate the public about the energy that ICT devices consume, the carbon dioxide that is produced from the use of ICT devices and how to reduce the energy and carbon dioxide emission from the use of ICT. Awareness of the new technologies that can help reduce both the energy consumption and carbon dioxide emission needs to be promoted. Energy harvesting is a developing research area and if it is to be taken up by society, it is important to educate society about the benefits of the new technology.

Acknowledgements

This work has been funded by EU Framework Programme 7 project ZEROPOWER http://www.zero-power.eu/.

Author details

Gabriel Abadal Berini[1], Giorgos Fagas[2], Luca Gammaitoni[3] and Douglas Paul[4]

1 Dept. d’Enginyeria Electrònica Escola d'Enginyeria, Universitat Autònoma de Barcelona, Barcelona, Spain

2 Tyndall National Institute, University College Cork, Cork, Ireland

3 NiPS Laboratory, Università di Perugia, Italy

4 James Watt Nanofabrication Centre, School of Engineering, University of Glasgow, Glasgow, UK

References

[1] SMART 2020: enabling the low carbon economy in the information age” is a report published by “The Climate Group”, an independent, not-for-profit organization. The report is available here: http://www.theclimategroup.org/_assets/files/Smart2020Report.pdf

[2] Disruptive Solutions for Energy Efficient ICT, Expert consultation, FET Proactive, 2010. Available at: http://cordis.europa.eu/fp7/ict/fet-proactive/docs/shapefetip-wp2011-12-10_en.pdf

[3] ITRS (International Technology Roadmap for Semiconductors), Semiconductor Industry Association, 2001, http://public.itrs.net

[4] R. K. Cavin, V.V. Zhirnov, D. J. C. Herr, A. Avila, and J. Hutchby, Research directions and challenges in nanoelectronics, J. Nanoparticle Res., vol. 8, pp. 841–858, 2006

[5] G. Baccarani, M. R. Wordeman and R. H. Dennard, IEEE Trans Electron Devices, 31(4), 452–62, 1984

[6] Energy Harvesting and Storage for Electronic Devices 2011-2012 – IDTechEx (http://www.idtechex.com/research/reports/energy-harvesting-and-storage-for-electronic-devices-2011-2021-000270.asp)

[7] Energy Harvesting for Structural Monitoring – A Roadmap to New Research Challenges, UK EPSRC EH Network Workshop Report, May 2011 (http://eh-network.org/resource1.php)

[8] Energy Harvesting from Human Power – A Roadmap to New Research Challenges, UK EPSRC EH Network Workshop Report, March 2011 (http://eh-network.org/resource1.php)

[9] Power Management Technologies to Enable Remote and Wireless Sensing, UK ESP KTN Report, May 2010 (http://eh-network.org/resource1.php)

[10] Energy Harvesting Technologies to enable Wireless and Remote Sensing, UK Sensors and Instrumentation KTN Report, June 2008, (http://eh-network.org/resource1.php)

Energy Management at the Nanoscale

Luca Gammaitoni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57345

1. Introduction

Energy management is considered a task of strategic importance in contemporary society. It is a common fact that the most successful economies of the planet are the economies that can transform and use large quantities of energy. In this chapter we will discuss the role of energy with specific attention to the processes that happens at micro and nanoscale, that are the scales where the modern ICT devices are built and operated.

2. Energy and its transformation

We start our journey toward the role of energy in ICT devices by addressing the most fundamental question of what energy is. According to Richard Feynman[1] we do not know much about energy apart from the fact that "energy is conserved". We usually define energy as the "capability of performing work" and define work as the activity performed by a force applied to some mass that, as a consequence of this application, changes its position. Energy, in the International System of Units, is measured in Joule (symbol J). 1 J is equal to the work done by applying a force of one Newton through a distance of one metre. 1 J is the energy that it takes to raise an apple (100 g) for 1 m above the ground. It is also relevant how much time it takes to do some work. This is taken into account by the concept of *power*. Given a certain amount of work done, the power is equal to the work divided by the time that it takes to do it. Physical unit of power is Watt (symbol W; 1W = 1J /1s). Multiples or sub multiples of J or W are what we are used to deal with in our everyday life: one kilowatt or kW is equal to 1000 W and it is the power required to cook a cake in a microwave oven. Given that the cake takes about one

[1] "It is important to realize that in Physics today, we have no knowledge of what energy is". The Feynman Lectures on Physics Volume I, 4-1.

hour to be properly cooked, the amount of energy used at this aim is approximately 3600 x 1000 = 3.6 million J or 1 kWh (*kilowatt-hour*).

In the following table we list few examples of how much power is required to perform some common tasks.

Task	Power (W)
Average power of a Boing 747 airplane	10^8
Full power aircraft fighter	10^6
Full power car engine	10^5
Operate a microwave oven	10^3
Being alive for an average adult human	10^2
Brain functioning for an average human	10
mobile phone calling	1
Emission of a standard WI-FI router	10^{-1}
Functioning of a LED light	10^{-2}
Functioning of a miniature FM receiver	10^{-3}
Functioning o a wireless sensor node	10^{-4}
Low power radio module	10^{-5}
Functioning of a quartz wristwatch	10^{-6}
Operation of a quartz oscillator	10^{-7}
Sleep mode of a microcontroller	10^{-8}
1 bit information erasure at room T (min)	10^{-21}

Table 1. Order of magnitude of the power required to perform some common activities.

From our everyday experience we know that energy is used for many different tasks and comes in many different forms. We know that it is energy because we can use it to perform work, like moving a car or a train. Thus energy is a property of physical systems that can be used to perform work and usually comes inside physical objects like a hot gas or a gasoline tank. Thinking about it we can ask questions like: how can we make the energy contained in a litre of gasoline to push forward a car or how can we use the heat produced by burning coal to make the train run?

Questions like these were at the very base of the activities performed in the early seventeen hundreds by the first inventors of the so-called thermal machines. People like Thomas Newcomen (1664-1729) who built the first practical steam engine for pumping water and James Watt (1736-1819) who few decades after proposed an improved version of the same machine. It is thanks to the work of scientists like Sadi Carnot (1796-1832) and subsequently of Émile

Clapeyron (1799 - 1864), Rudolf Clausius (1822 - 1888) and William Thomson (Lord Kelvin) (1824 – 1907) that studies on the efficiency of these machines aimed at transforming heat (just a form of energy) into work brought us the notion of entropy and the laws of thermodynamics.

These laws do not tell us much about what energy is but they are very good in ruling what can we do and what we cannot do with energy. Let's briefly review them.

The first law of thermodynamics states that the total energy of an isolated physical system is conserved during any transformation the system can go through.

It was initially formulated by Julius Robert von Mayer (1814 - 1878) and subsequently reviewed by James Prescot Joule (1818-1889) and Hermann Ludwig Ferdinand von Helmholtz[2] (1821-1894). It is strongly believed to be true but, to some extent is a self-supporting law: as a matter of fact it is so strongly believed that in every instance we observe a possible violation we think harder to discover some way in which energy can be hidden and overlooked. Last time in history this happened was at the beginning of 1900 when Albert Einstein proposed the mass-energy equivalence to account for the "missing mass" during a nuclear transformation.

The second law states that there are limitations to how much work we can get from a given amount of energy present in the form of heat. There exist different formulations that are all equivalent. The two most popular are ascribed to Clausius and Kelvin:

Clausius formulation: "No process is possible whose sole result is the transfer of heat from a body of lower temperature to a body of higher temperature".

Kelvin formulation: "No process is possible in which the sole result is the absorption of heat from a reservoir and its complete conversion into work".

An important consequence of the second law is that there is a limit to the efficiency of a thermal machine. This limit was discovered by Sadi Carnot in 1824 when he was only 28. In the publication entitled *Réflexions sur la Puissance Motrice du Feu* ("Reflections on the Motive Power of Fire") he introduced the concept of thermal machine, generalizing the concept popular at that time of "steam engine", and showing that the efficiency of any thermal machine operating between two temperatures is bounded by a quantity that is a function of the two temperatures only.

Few years after the work of Carnot, Clausius used this result to introduce a quantity that is useful in describing how much heat can be changed into work during a transformation. He proposed the name "entropy" for his quantity. The idea is the following: if you want to operate a thermal machine you have to find a cyclic transformation during which heat is changed into work. The cycle is necessary because you want to operate the machine continuously and not just once. Clausius proved a theorem that states that during a cyclic transformation, if you do the transformation carefully enough not to loose any energy in other ways (like friction), then

[2] Feynman assertion that the notion of energy has never been very clear is testified by the fact that the key publication of Helmholtz, considered the father of the conservation of energy, is entitled "Über die Erhaltung der Kraft", "On the conservation of the strength". On the other hand the kinetic energy has been called for a long time "vis viva", Latin expression for "living strength".

the sum of the heat exchanged with the external divided by the temperature at which the exchange occurs is zero:

$$\oint \frac{dQ}{T} = 0 \tag{1}$$

The cycle does not depend on the path that you take and, clearly you start and end at the same state. This is equivalent to say that it exists a state function S defined as

$$S_B - S_A = \int_A^B \frac{dQ}{T} \tag{2}$$

(or in differential form $dS = dQ/T$) that satisfies the previous equation. If you are not careful enough and you loose energy during the transformation than the inequality holds instead:

$$\oint \frac{dQ}{T} \leq 0 \tag{3}$$

A transformation like this is also called an *irreversible transformation*. It is easy to show that if we take and irreversible transformation to compute the entropy we end up with under-estimating the change:

$$S_B - S_A \geq \int_{A\ irr}^{B} \frac{dQ}{T} \tag{4}$$

In the particular case in which we are considering a transformation without any heat ex-changed then the second term is zero and the final entropy is always larger than the initial one. A typical example is the so-called adiabatic expansion of a gas. If we consider an infinitesimal transformation we have:

$$dS \geq \frac{dQ}{T} \quad \text{or} \quad TdS \geq dQ \tag{5}$$

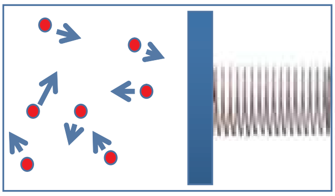
where the equal sign hold during a reversible transformation only. The previous equation is sometimes considered a concise formulation of the second principle of thermodynamics. If I put in contact a physical system that is at temperature T1 with a heat reservoir that is at temperature T2>T1 then some heat is transferred from the reservoir to the system. Accordingly the integral is positive and the entropy of the system increases. The other way around phenomenon by which heat is transferred from the system to the reservoir does not happen (second principle) and thus we conclude that during a spontaneous transformation (i.e. without external work) the entropy always increases. We can make the entropy of our system decrease (like in a refrigerator) but we have to add work from outside.

Back to the Clausius inequality, it is useful to interpret the quantity TdS as the amount of heat (meaning thermal energy) that cannot be used to produce work. In other words during a transformation, even if we are carefully enough not to waste energy in other ways, we cannot use all the energy we have to do useful work. Part of this energy will go into producing an increase of the system entropy. If we are not carefully enough the situation is even worst and we get even less work. This is sometimes accounted by the introduction of the so-called *Free energy*. The concept of Free energy was proposed by Helmholtz in the form: $F = U - TS$. The free energy F measures the maximum amount of energy that we can use when we have available the internal energy U of a system.

In summary we can state that change in entropy is a measure of a system's thermal energy per unit temperature that is unavailable for doing useful work.

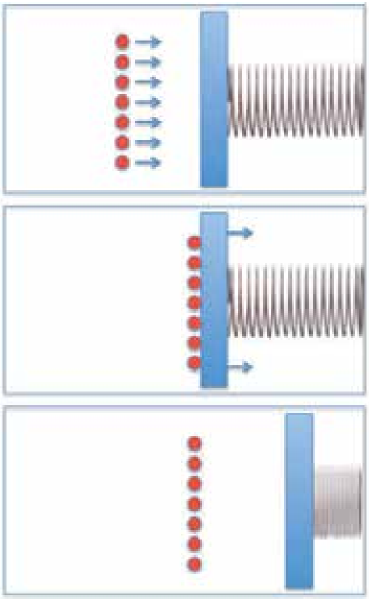
3. The microscopic perspective

Notwithstanding the above explanation, available since the late 1800, in general the entropy remained an obscure quantity whose physical sense was (and somehow still is) difficult to grasp. It was the work of Ludwing Boltzmann (1844 – 1906) that shed some light on the microscopic interpretation of the second law (and thus the entropy) as a consequence of the tendency of a physical system to attain the equilibrium condition identified with the most probable state among all the possible states it can be in. The ideal world of Boltzmann is made by physical systems constituted by many small parts represented by colliding small spheres[3]. To better grasp the meaning of entropy let's consider an ideal gas made by N particles in the form of tiny hard spheres of mass m that can collide elastically, i.e. conserving the kinetic energy in addition to their momentum. Let's suppose that these particles are contained in a box that has a moving set of mass $M = Nm$. The set is connected to a spring of elastic constant k, as in the figure, and is at rest.



If all the particles have the same velocity v and collide perpendicularly with the moving set at the same time (see following figure), they will exchange velocity with the set.

[3] L. Boltzmann was a strong supporter of the atomistic view of the matter. An idea that was not given for granted (nor so popular) at his time. This did not help an often depressed Boltzmann that in 1906, during one of his bad mood crises decided to kill himself.



This will compress the spring up to an extent x_1 such that:

$$\frac{1}{2}M\ v^2 = \frac{1}{2}k\ x_1^2 = U \qquad (6)$$

This is a simple transformation of kinetic energy into potential energy. We can always recover the potential energy U when we desire and use it to perform work. The work will be exactly U. In this case we can completely transform the energy of the gas particle into work. How comes? Well, in this case we are clearly considering a very special configuration of our gas. Unique indeed. What is on the contrary the most probable configuration for the particle in the gas? Based on our experience (and on some common sense as well) it is the configuration in which all the particles, although each with the same velocity v, are moving with random direction in the box (as in our first figure). The energy of the gas is still the same (so is its temperature T) but in this case the set will be subjected at random motion with an average compression of the spring such that its average energy is U/N. This is also the maximum work that we can recover from the potential energy of the movable set. Thus it appears clear that, although the total energy U is the same in the two cases, in the second case we have no hope of using the greatest part of this energy to perform useful work. As we have learned above, when we introduced the definition of Free energy, the quantity that limits our capability of performing work is the entropy. Thus the systems that have the smaller entropy have the larger capability of performing work. Accordingly we can use the entropy to put a label on the energetic content of a system. Two systems may have the same energy but the system that has the lower entropy will have the “most useful” energy.

This example helped us to understand how energy and entropy are connected to the microscopic properties of the physical systems. In the simple case of an ideal gas, the system energy is nothing else than the sum of all the kinetic energies of the single particles. We can say that the energy is associated with "how much" the particles move. On the other hand we have seen that there is also a "quality" of the motion of the particles that is relevant for the entropy. We can say that the entropy is associated with "the way" the particles moves. This concept of "way of moving" was made clear by Boltzmann at the end of 1800, who proposed for the entropy the following definition:

$$S = K_B \log W \tag{7}$$

where K_B is the famous Boltzmann constant and W is the number of microstates associated with a given state of the physical system. Sometimes W is also called the "number of configurations" and represents the number of ways we can arrange all the particles in the system without changing its macroscopic properties[4]. In the previous example we have only one way to arrange the N particles so that they are all parallel, aligned and with the same velocity while we have a very large number of ways of arranging the N particles to be a randomly oriented set of particles with velocity v. Thus it is clear that in the second case the value of the entropy is much larger than in the first case (where it is indeed zero).

We have seen above that during a spontaneous transformation the entropy of the system increases. This can happen without any change in the energy of the system itself[5]. In our example with the bouncing particles this is represented by the situation in which the particles trajectories are not perfectly aligned as a consequence collision between the particles can happen before hitting the movable set and soon, collision after collision the entire group of particles evolves into a randomly moving group. This is clearly a spontaneous transformation. By the moment that the collision are elastic the energy of the system has not changed but the system entropy has rapidly increased up to its maximum value. Conversely the free energy has reached its minimum value. This is what we call a "spontaneous transformation toward the equilibrium condition". Now: can we bring the system back to its initial condition? The answer is yes but… in order to do it we need to spend some energy as required by the second principle. How much? Clearly we need to spend $T\Delta S$ of energy, where ΔS represents the difference in entropy between the final and the initial state. The bad news is that if we spend this energy and decrease the entropy back to its original condition… the energy is lost and we cannot recover it any more, meaning that the energy that we spend does not change the total kinetic energy of the system that remains the same. In other words there is no way to store energy into decreasing entropy. This last conclusion has some consequences that we will explore soon below when we will consider the energy processes in ICT devices. Before moving

[4] Here we assume that all the microstate are equiprobable. The extension to the more general case with microstates with different probabilities has been proposed by Josiah Willard Gibbs (1839 – 1903).

[5] There is a famous experiment performed by Joule that shows that the free expansion of a gas is a process that can happen without exchange of energy.

into that subject however we need to dig some more into the consequences of thermodynamics in the small scale[6].

4. What does irreversible mean?

When we introduced the entropy change we specified that this is defined in terms of heat transfer, once we perform a *reversible transformation*. What does it exactly mean reversible? Well, reversible literally means that "it can be done the other way around" but in my opinion it is not a very clear definition. What is usually meant is that if we want to go from a state A toward a state B, we do need to do a transformation that it is so slow that it goes through an infinite number of equilibrium states so that at any instant all the macroscopic quantities like temperature, pressure, volume, … are well defined. By the moment that these quantities are defined only in equilibrium condition we need to be as close to equilibrium as possible. For a number of comprehensible reasons that we will address more in detail in another chapter, this requires that we go quite slow[7] when we change anything in the system.

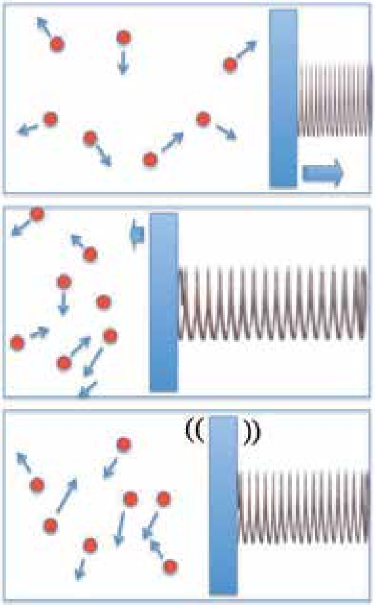
What happens if we do not go "slow"? Well, as we have seen before, in this case we are performing an *irreversible transformation*. During an irreversible transformation the entropy always increases. Moreover, due to the Clausius **inequality** it always increases some more compared to what it would be required by the second law. Why is that? The answer is that in addition to the *physiological* increase there is an extra contribution due to the *dissipative effect* of the non-equilibrium processes. With *dissipative effect* we intend a way in which some low-entropy energy is changed into high-entropy energy. A typical example of dissipative process is friction. If during any transformation there is friction then the transformation is irreversible and the increase in entropy *benefits* from the additional contribution of this process.

In this regards it is interesting to inspect more in detail the example of the movable set in contact with the gas, we introduced before. When the system represented by the particle gas + the movable set is at equilibrium the movable set is not only acted on by the collision of the particles but is also damped by the very same source. To see this effect we can consider two simple cases.

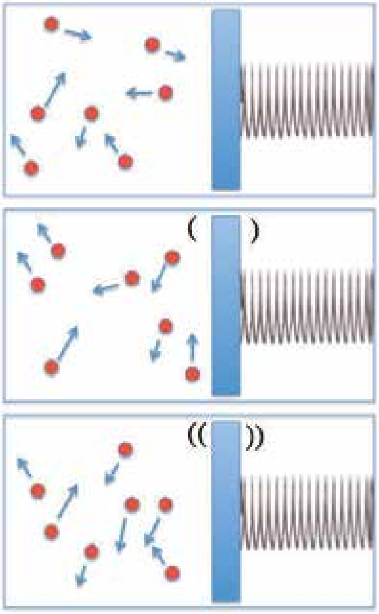
1. We compress the spring to some extent and then we release the compression leaving it free to oscillate. After few oscillations we observe that the oscillation amplitude decreases as a consequence of what we call the friction (viscous damping force) action due to the presence of the gas. The decrease ceases when the oscillation amplitude reaches a certain equilibrium value and after that it remains constant (on average, see following figure). Some energy has been dissipated into heat.

[6] Thermodynamics in the small scale is a kind of oxymoron. According to J. P. Sethna, thermodynamics is the theory that emerges from statistical mechanics in the limit of large systems (J.P. Sethna, Statistical Mechanics: Entropy, Order Parameters and Complexity, 6.4, Oxford Univ. Press. 2008).

[7] Here slow means slow compared to the time it take for the system to relax to equilibrium.



2. We now start with the movable set at rest and leave it free. After few seconds we will see that the set starts to move with increasing oscillation amplitude that soon reaches an equilibrium condition at the very same value (on average) of the first case (see following figure).



In both cases the two different roles of damping-force and pushing-force has been played by the gas. This fact led to think that there must be a connection between the process of dissipating energy (a typical irreversible, i.e. non-equilibrium process) and the process of fluctuating at equilibrium with the gas.

5. A bridge toward non-equilibrium: fluctuation-dissipation relation

In order to unveil such a link we need to introduce a more formal description of the dynamics of the movable set. This problem has been addressed and solved by Albert Einstein (1879 - 1955) in his 1905 discussion of the Brownian motion and subsequently by Paul Langevin (1872 - 1946) who proposed the following equation:

$$m\ddot{x} = -m\gamma\dot{x} - \frac{dU}{dx} + \xi(t) \tag{8}$$

As before x represents the movable set position. Here γ represents the viscous damping constant, U is the elastic potential energy due to the spring and $\xi(t)$ is the random force that accounts for the incessant impact of the gas particles on the set, assumed with zero mean, Gaussian distributed and with a flat spectrum or, delta-correlated in time (white noise assumption):

$$\langle \xi(t_1)\xi(t_2)\rangle = 2\pi\, G_R \delta(t_1 - t_2) \tag{9}$$

where the <> indicates average over the statistical ensemble.

Now, as we noticed before, by the moment that the gas is responsible at the same time for the fluctuating part of the dynamics (i.e. the random force $\xi(t)$) and the dissipative part (i.e. the damping constant γ) there must be a relation between these two. This relation has been established within the linear response theory (that satisfies the equipartition of the energy among all the degrees of freedom) initially by Harry Theodor Nyquist (1889 - 1976) in 1928[8], and demonstrated by Callen and Welton in 1951. This relation is:

$$G_R = \frac{mK_BT}{\pi}\gamma \tag{10}$$

and represents a formulation of the so-called Fluctuation-Dissipation Theorem (FDT)[1,2]. There exist different formulations of the FDT. As an example we mention that it can be generalized to account for a different kind of dissipative force, i.e. internal friction type where γ is not a simple constant but shows time dependence (work done in the sixties by Mori and Kubo). In that case the random force shows a spectrum that is not flat anymore (non-white noise assumption).

[8] Nyquist established this relation while he was studying the voltage fluctuations across an electrical resistor. The random electromagnetic force arising across a resistance at finite temperature is a function of the value of the resistance itself.

http://dx.doi.org/10.5772/57345

Why is FDT important? It is important because it represent an ideal bridge that connects the equilibrium properties of our thermodynamic system (represented by the amplitude and character of the fluctuations) with the non-equilibrium properties (represented here by the dissipative phenomena due to the presence of the friction). Thus there are basically two ways of using the FDT: it can be used to predict the characteristics of the fluctuation or the noise intrinsic to the system from the known characteristics of the dissipative properties or it can be used to predict what kind of dissipation we should expect if we know the equilibrium fluctuation properties. Its importance however goes beyond the practical utility. Indeed it shows like dissipative properties, meaning the capacity to produce entropy, are intrinsically connected to the equilibrium fluctuations.

6. Energy transformations for small systems

How does the description of the energy transformation processes presented so far change when we deal with small systems? To answer this question we start considering an important aspect when we deal with physical systems: the condition of being an *isolated system*. If we say that a system is not isolated, we intend that it has interactions of some kind with something that we consider external to the systems itself. If this is not the case (isolated system) all the dynamics is self-determined by the system itself and we can deal with it by addressing the equations of motion for each particle coupled to each other particle in the system. At this aim we may use the standard Newton laws (or in the quantum case the Schrodinger equation). If the system is *not isolated* the situation is generally more complex and we need to take into account the interaction of our system with the "external world". In principle however any system can be considered isolated provided that we include in the system all the sources of interactions. In the extreme case we can consider the universe itself as an isolated system. For this reason we will limit our consideration to systems that are isolated.

Before answering the question about the energy transformations in small systems we should be more precise in defining what a *small system* is. When we deal with real physical systems we cannot ignore that all the matter, as we know it, is composed by atoms. These are more or less individual particles whose interactions determine most of the properties that characterize the matter. The ordinary devices that we are used to deal with are composed by very a large assembly of atoms, numbers are of the order of the Avogadro number, i.e. NA = 6.022×10^{23}. Thus when we are dealing with small systems, in general we intend systems composed by a number of atoms N that is small compare to NA. Clearly, due to the extremely large value of NA, a system composed by few thousands of atoms (or molecules or "particles") can still be considered small. This is the case for example of the nanodevices like last generation transistors. Unfortunately in this case the small systems are not isolated because they exchange energy and information with the outside. On the other hand small isolated systems are quite rare. An example of the *small isolated system* can be found in the realm of what is generally called "high energy physics": here the particles are most of the time just few (small system) and isolated from the external. Back to the realm of the physics of matter we have frequently to deal with systems that are usually not small but can be considered in good approximation isolated. What do we do in these cases?

One possibility is to do what we did just before, when we dealt with the movable set in contact with the gas of N particles. Here N is of the order of N_A. Overall our system is composed by 3N+1 degrees of freedom (dof): 3 for each of the N particles and 1 for the movable set position coordinate x. This is clearly not a small system, although isolated, because all the interactions are inside the 3N+1 dof. In this case we played a trick: we focused our attention on the single degree x and summarized the role of the remaining 3N dof by introducing the dissipative and the fluctuating force like external forces. By the moment that both these forces are necessary to account for the observed dynamics and by the moment that both are born out of the neglected 3N dof, it comes out that they are connected to each other and the connection is nothing else that the FDT that we discussed. Our equation of motion is not anymore the deterministic Newton (Schrodinger) equation, instead it is the novel stochastic Langevin equation where there is a both friction and fluctuation caused by the added forces due to the neglected dof. Thus the trick we played was to exchange the dynamics of a *not small isolated system* with *small not isolated system.* Such an approach has different names (adiabatic elimination, coarse graining, …) and it is considered a very useful tool in describing the properties of dynamical systems composed by many dof.

To summarize our approach: we have transformed a *non small isolated* system into a *small non isolated system*. What is the advantage? Easy to say: the dynamics of a non small isolated system can be described in terms of 3N+1 dof by 3N+1 coupled motion equations and when N is of the order of N_A this is a practically impossible approach. Thus the advantage was to drastically reduce the number of equations of motion (in this case to just 1) but the price we had to pay is the introduction of dissipation and fluctuation. What we have found is that dissipative and fluctuating effects appear only if we neglect some (usually many) dof through some coarse graining approximation to the system dynamics. In this perspective the dissipation of energy appears to be only an illusion due to our choice of dynamical description.

On the other hand we know that if we perform a real experiment with our movable set, indeed we observe a decrease in the oscillation amplitude of the set until it reaches the stop and then it does start to fluctuate around the equilibrium position. This is not and illusion. The potential energy initially stored in the spring is now dissipated due to the presence of the gas particles. How does this fit with what we just said about the dissipation being an illusion? The answer is that the total energy (the kinetic energy of the gas particles + the potential energy initially stored in the spring) is conserved because the (not small) system is isolated. What happened is that the potential energy of the movable set has been progressively transformed into additional kinetic energy of the N particles that now have a slightly larger average velocity (the temperature of our gas slightly increased). Thus during the dynamics the energy is transferred from some (few) dof to others (many) dof. This is what we called before energy dissipation and now it appears to be nothing more than energy re-distribution. Before we have seen that dissipative effects during a transformation are associated with an increase of entropy. Indeed this energy distribution process is an aspect of the tendency of the system to reach the maximum entropy (while conserving the energy). This is what we have called a spontaneous transformation: the increase of the entropy up to the point where no more energy distribution process takes place, i.e. the thermal equilibrium.

Is this the end of the story? Actually it is not. There is a quite subtle aspect that is associated with the conservation of energy. It is known as Poincaré recurrence theorem. It states that in a system that conserves energy the dynamics evolve in such a way that, after a sufficiently long time, it returns to a state arbitrarily close to the initial state. The time that we have to wait in order to have this recurrence is called the Poincaré recurrence time. In simple words this means that not only the dissipation of energy is an illusion because the energy is simply redistributed among all the dof but also that this redistribution is not final (i.e. on the long term the equilibrium does not exist). If we wait long enough we will see that after some time the energy will flow back to its initial distribution and our movable set will get its potential energy back (with the gas particle becoming slightly colder). This is quite surprising indeed because it implies that in this way we can reverse the entropy increase typical of processes with friction and thus fail the second principle. Although this may appear a paradox this answer was already included in the description of entropy proposed by Boltzmann and specifically in its intrinsic probabilistic character. The decrease of entropy for a system composed by many dof is not impossible: it is simply extremely improbable. It goes like this: for any finite observation time the dynamic system evolves most probably in a direction where the entropy increases because according to Boltzmann this is the most probable evolution. However if we wait long enough also the less probable outcome will be realized and thus the second principle *violated*. How much time should we wait? The answer depends on the dof of our isolated (energy conserving) system. The larger the number of dof the longer the time to wait… exponentially longer.

Author details

Luca Gammaitoni*

NiPS Laboratory, Università di Perugia, Italy

References

[1] Nyquist H (1928). "Thermal Agitation of Electric Charge in Conductors". Physical Review 32: 110–113.

[2] H. B. Callen, T. A. Welton (1951). Physical Review 83: 34.

[3] L. D. Landau, E. M. Lifshitz. Physique Statistique. Cours de physique théorique. 5. Mir.

[4] J.P. Sethna, Statistical Mechanics: Entropy, Order Parameters and Complexity, 6.4, Oxford Univ. Press. 2008

[5] Umberto Marini Bettolo Marconi; Andrea Puglisi; Lamberto Rondoni; Angelo Vulpiani (2008). "Fluctuation-Dissipation: Response Theory in Statistical Physics". Physics Reports 461 (4–6): 111–195.

[6] Macroscopic equations for the adiabatic piston, Massimo Cencini, Luigi Palatella, Simone Pigolotti, Angelo Vulpiani. Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), Vol. 76, No. 5. (2007).

Chapter 3

Kinetic Energy Harvesting

Helios Vocca and Francesco Cottone

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57091

1. Introduction

The recovery of wasted energy present in the ambient that is a reject of artificial or natural processes to power wireless electronics is paving the way for enabling a huge number of applications. One of the main targeted technologies that meets the levels of harvestable power, typically few hundreds of microwatts, is represented by wireless sensor networks (WSNs) [1]. This technology consists of a grid of spatially-distributed wireless nodes that sense and communicate information like acceleration, temperature, pressure, toxicity of the air, biological parameters, magnetic field, light intensity and so on, among each other and up to the end user through a fixed server. In the next years, WSNs will be massively employed in a wide range of applications such as structural monitoring, industrial sensing, remote healthcare, military equipment, surveillance, logistic tracking and automotive monitoring. In fact, harvesting energy directly from the ambient not only represents a realistic mean to integrate or substitute batteries, but is the sole way for enabling many contemporary and future wireless applications that will be all integrated in the so called "internet of things" [2].

Actually, WSNs already have the characteristics of ubiquity, self-organizing and self-healing but they would not be deployable unless they will also be self-powering. As a matter of fact, it is very expensive and impractical to change batteries in most of the anticipated potential applications. For long-term operation in inaccessible or harsh locations, energy harvesting is a key solution. For example, long-term environmental, structural health of buildings or bridge monitoring and control would require many thousands of integrated sensors impossible to be replaced or maintained. The possibility for chronically ill patients to be continuously monitored without changing batteries would represent a significant improvement in their life quality.

Among various renewable energy present in the environment such as solar, radio frequency RF, temperature difference and biochemical, kinetic energy in the form of mechanical vibrations is deemed to be the most attractive, in the low-power electronic domain, for its power density, versatility and abundance [3]. This type of energy source is located in buildings, vibrating machineries, transportations, ocean waves and human beings, and it can be converted to power mobile devices.

The power consumption of wireless sensors has been largely reduced in the last years thanks to the Ultra-Low-Power electronics [4]. Typical power needs of mobile devices can range from few microwatts for wristwatches, RFID, MEMS sensors and actuators up to hundreds of milliwatts for MP3, mobile phone and GPS applications. They are usually in a sleep state for the 99.9% of their operation time, waking up for a few milliseconds only to communicate data. Consequently, their average power consumption has been reduced below 10μW in order to match the power density capability of current generators (100-300 microwatts per cubic centimeter). For comparison, a lithium battery can provide 30μW/cc for 1 year or 30mW/cc for just 10 hours, while a vibration-driven generator could last for at least 50 years with the same power level [5]. Along with virtually infinite operational life, many other benefits come from motion-driven energy harvesting: no chemical disposal, zero wiring cost, maintenance-free, no charging points, capability for deployment in dangerous and inaccessible sites, low cost of retrofitting, inherent safety and high reliability.

A typical integrated vibration-powered wireless sensor includes an embedded vibration energy harvester (VEH), multiple-sensor module, microcontroller and a transceiver (Figure 1). Due to the variable nature of vibrations in their intensity and frequency, the device also contains an AC/DC voltage regulation circuit, which in turn can recharge a temporary storage system, typically a super-capacitor or a thin film Lithium battery. Capacitors are usually preferred as temporary storage systems for their longer lifetime, higher power density and fast recharging. In some applications, however, a storage system is not even necessary. The vibration energy harvester module is often tailored for the specific application and vibration spectrum of the source: harmonic excitation, random noise or pulsed movement.

2. Main conversion techniques

There are three main categories of kinetic-to-electrical energy conversion systems: piezoelectric, electrostatic and electromagnetic. In addition, there is the magnetostrictive branch as a variant of piezoelectric except for the use of magnetically polarized materials [6]. Each technique presents advantages and drawbacks. Therefore, there not exist a technique suitable for all cases and the optimal choice depends on the specific application.

Piezoelectric transducers make use of electrically polarized materials such as Barium Titanate (BaTiO3), Zinc Oxide (ZnO) and Lead Zirconate Titanate ($Pb[Zr_xTi_{1-x}]O_3$), commonly known as PZT which is considered one of the best materials for high electromechanical coupling. The direct piezoelectric effect used for energy harvesting was early discovered by French physicists Jacques and Pierre Curie in 1880. It occurs when an electric charge is generated within a material in response to applied mechanical stress (Figure 2). The strain and coupling coefficients in the fundamental piezoelectric equations are in general much higher in 33 mode than

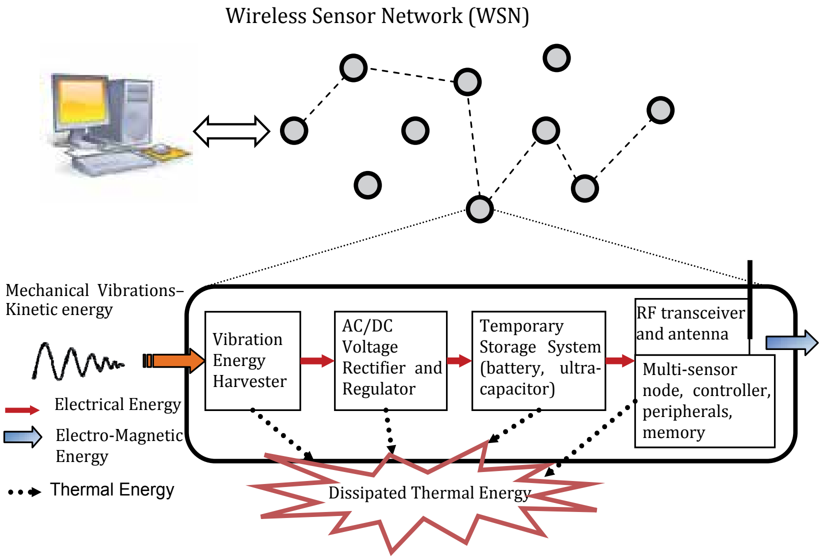


Figure 1. Wireless sensor network and vibration-driven wireless node with power fluxes.

in 31 [7]. However the 33 mode of bulk crystal corresponds to very high natural frequencies (~1 to 100 kHz), while longitudinal strain is easily produced within a cantilever beam that resonates at lower frequencies (~100 Hz) (Figure 2c).

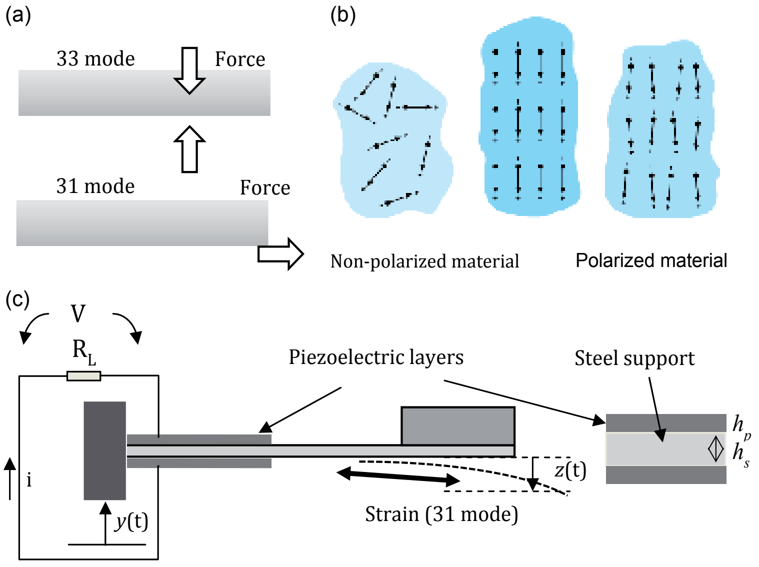


Figure 2. (a) Direct piezoelectric effect with 33 and 31 strain-charge coupling. (b) Polarization process scheme. (c) Drawing of bimorph piezoelectric cantilever beam.

Piezoelectric systems are capable of high voltage level (from 2 to several volts), well adapted for compact size and very good in terms power density per unit of volume. However, piezoelectric coupling decreases very fast at micrometric scale and relatively large load impedances are required to reach the optimal working point [8]. Besides, other problems must be considered such as aging, depolarization and brittleness. For low frequency applications, like those related to wearable sensors, polymer-based materials (e.g. dielectric elastomers) constitute a valid alternative to ceramics because of their flexibility, inexpensiveness and durability [9].

Electromagnetic technique is simply realized, according to Faraday's law, by coupling a static magnetic field produced by a permanent magnet and a solenoid in relative motion – one of which usually acts as a stator; the other as a mover. These systems show complementary behaviour in terms of frequency bandwidth and optimal load in relation to piezoelectric techniques. They are recommended for lower frequencies (2-20 Hz), small impedance and medium size [10]. Furthermore, their cost is smaller than other solutions. Most of the commercial solutions are available at centimetre scales because they exhibit higher power density than piezoelectric devices. On the other hand, the integration of electromagnetic harvesters into micro-electro-mechanical-systems (MEMS) results difficult. However, some of these limitations have been overcome to date [11-13].

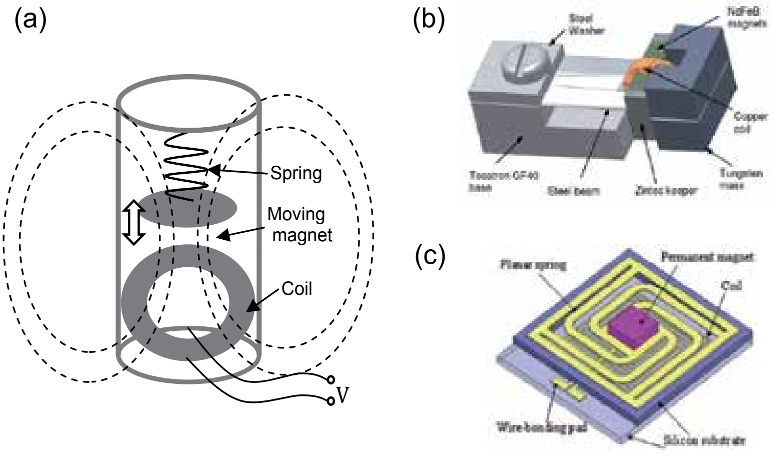


Figure 3. (a) Simple architecture of em-VEH where a moving magnet (mover) oscillates with respect to a fixed coil (stator). (b) Moving magnet across coil arrangement with discrete components [11-13]. (c) Microfabricated em-VEH where a small magnet oscillates towards a planar coil [11].

Electrostatic harvesters basically consist of a variable capacitor in which one electrode is attached to an oscillating mass suspended by beams and the counter electrode is fixed elsewhere in the support structure. When a force is applied to the mass either the dielectric gap or the overlap surface of electrodes varies depending on the moving direction: the first case being referred to as an in-plane gap-closing converter (Figure 4a), while the second case as an overlap varying converter (Figure 4b). As a consequence, the capacitance changes and additional charges occur at the electrodes in order to balance the bias voltage. Hence, during the movement of the proof mass, a current flows through a load shunted to plates. A similar method of fixed bias voltage is that of charge constrained where a constant charge is held into the plates by means of a battery or another capacitor.

One of the main disadvantages of electrostatic vibration harvesters is the need of an external voltage source in order to be pre-charged. This fact seems to contrast with the goal of energy recycling from the ambient, but makes sense if the source comes from the energy storage associated to the harvester [12]. In that case the generator only needs to be kick-started at the beginning of the conversion process. Some designs overcome this problem by using electrets to provide the pre-charge bias voltage [14].

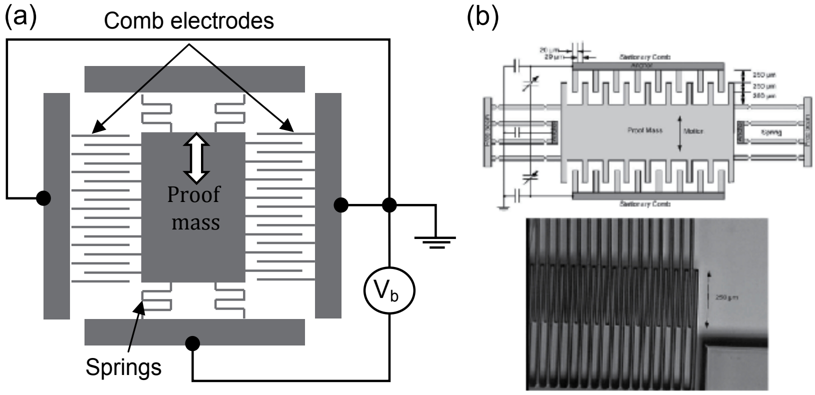


Figure 4. (a) Schematic of comb gap-closing electrostatic VEH. (b) Example of MEMS in-plane overlap electrostatic VEH [15].

Nevertheless, electrostatic technology is very well suited for MEMS manufacturing as employs the same elements of micro accelerometers [14]. Moreover, the silicon based MEMS do not have problems of aging as for piezoelectric materials. However the generated power is pretty much small compared to piezoelectric and electromagnetic [16].

Nowadays, vibration energy harvesters are delivered on the real market by some leading companies such as Perpetuum Ltd, Ferro Solution and Mide Volture although they have quite bulky size. The power density of commercial harvesters ranges from 10 to 300μW/cc relative to acceleration levels of 0.01-1g rms. Prototypes of MEMS-based harvesters have been demonstrated by universities and private teams, though they are still at experimental stage. Beeby et al. have implemented a vibration-powered wireless sensor node with embedded micro-electromagnetic generator [3]. Millimetre-sized electrostatic generators were formerly realized by Roundy et al. [11]. Miao implemented a parametric generator for biomedical applications [17] and examples of piezoelectric nano-mechanical generators are also emerging [18]. Mahmood and Basset have successfully built and tested an efficient MEMS-based electrostatic harvester [19, 20].

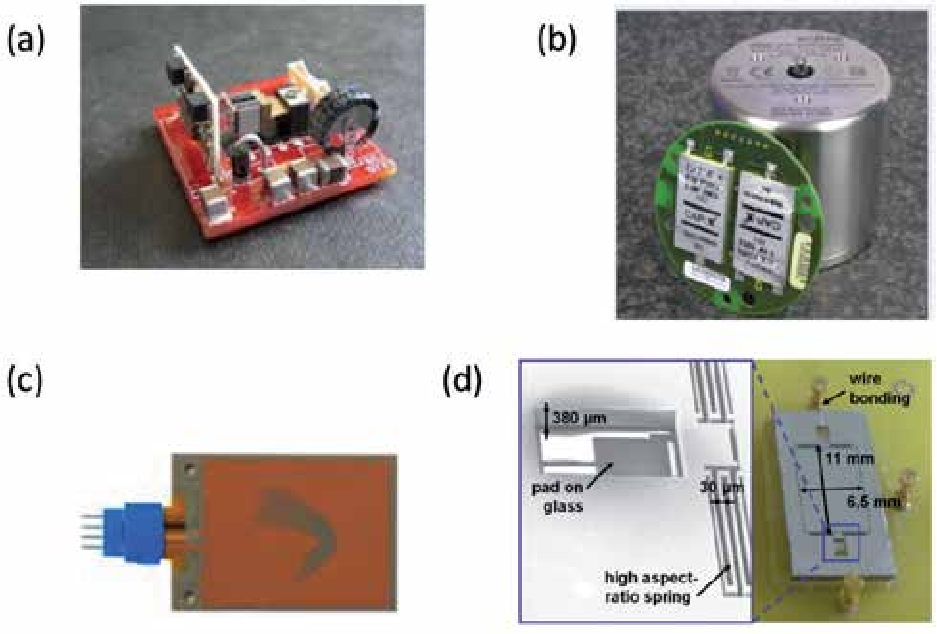


Figure 5. Examples of micro vibration energy harvester: (a) electromagentic [16] and (b) (Perpetuum), (c) piezoelectric (Midé), (d) electrostatic [11].

3. Linear spring-mass-damper models of VEHs

Kinetic energy harvesters are divided into two categories: those that utilize direct application of force and those that make use of the inertial force associated to a moving mass m. Inertial generators, are preferred to direct-force devices for vibration energy harvesting as they only need one point of attachment to a vibrating structure, thus allowing a simpler miniaturization.

Figure 6 illustrates basic models of (a) direct and (b) inertial force vibration-based generators independent from the conversion technology. In the second case, the driving force $F(t)$ is equal to $-m\ddot{y}$, where the base vibrations are represented as $y(t)$ and dot stands for the derivative with respect to time. $z(t)$ is the relative motion between the housing and proof mass, k is the spring stiffness, d is the parasitic damping, f_e represents the electrical restoring force due to the transduction mechanism. Finally, the electrical part includes the resistive load R_L through which flows the generated current i.

Williams and Yates early defined a basic technology independent model of micro-electric generator for vibration energy harvesting [16]. In that case the conversion force f_e being considered as an electrical damping force proportional to the velocity $f_e = -d_e\dot{z}$. However, the electrical restoring force can in general be a complex function of the mass displacement, velocity and acceleration. In addition, such an approximation does not take into account of the effect of the electrical branch coupled to the mechanical system.

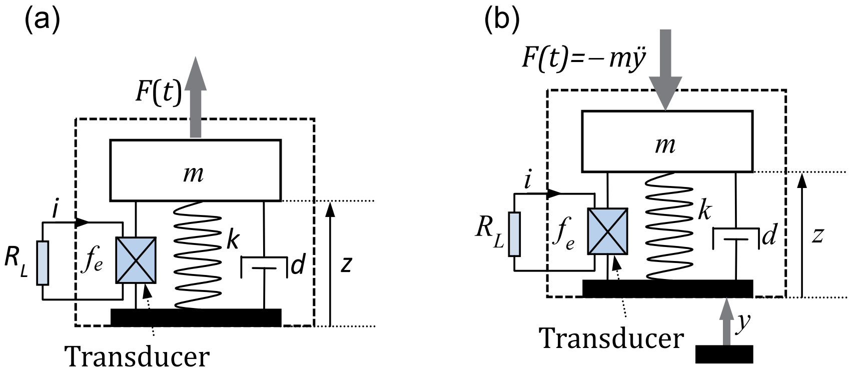


Figure 6. Models of non-technology specific (a) direct-force and (b) inertial-force generators. The electrical force f_e can be in general a complicated function of displacement, velocity or acceleration.

In the following we describe the lumped parameters model of the VEH by including the electrical domain as sketched in Figure 6b. This can be then applied to different types of linear conversion systems. The coupled governing equations of a generic 1-DOF vibration-driven generator are derived by the second Newton's law and Kirchhoff's law as follows

$$m\ddot{z} + d\dot{z} + kz + \alpha V = -m\ddot{y}, \qquad (1)$$

$$\dot{V} + (\omega_c + \omega_i)V = \omega_c \lambda \dot{z}, \qquad (2)$$

where dots stand for time-derivatives. The first equation describes the dynamics of the inertial mass while the second equation accounts for the coupled electrical circuit. V is the produced voltage across the electrical resistance and α is the electromechanical coupling factor; ω_c represents the characteristic cut-off frequencies of the electrical circuit of the system operating as high-pass filter due to the specific transduction technique. This parameter is the inverse to the characteristic time τ of the electrical branch, that is ω_c=1/τ. ω_i has the same meaning but corresponding to the internal resistance R_i of the system. Finally, λ is the electromechanical conversion factor. These characteristic parameters are derived from the harvester design depending on the specific conversion method and architecture as explained in the following examples. Hereafter we will only treat the piezoelectric and electromagnetic case, as the electrostatic is inherently nonlinear when considering the electrical force between close electrostatic plates.

3.1. Case 1: piezoelectric cantilever generator

Starting by the constitutive piezoelectric equations [21], for a bimorph piezoelectric cantilever like the one shown in Figure 2c with active layers wired in parallel, the above characteristic parameters are derived

$$\begin{array}{llll} \alpha = kd_{31} / h_p k_2, & \text{(a)} & \lambda = \alpha R_L, & \text{(b)} \\ \omega_c = 1 / R_L C_p, & \text{(c)} & \omega_i = 1 / R_i C_p, & \text{(d)} \end{array} \tag{3}$$

where h_P and h_s are the thickness of piezoelectric and support layer respectively; d_{31}, $E_{\rm p}$, ε_0, ε_r and C_p are the piezoelectric strain factor, Young's modulus, vacuum and relative dielectric permittivity and, finally, the equivalent capacitance of piezoelectric beam. These constants are related to the following structural parameters as derived in [7]

$$\begin{array}{llll} k = k_1 k_2 E_p, & \text{(a)} & & \\ k_1 = \dfrac{2I}{b(2l_b + l_m - l_e)}, & \text{(b)} & k_2 = \dfrac{3b(2l_b + l_m - l_e)}{l_b^2 \left(2l_b + \dfrac{3}{2} l_m \right)}, & \text{(c)} \\ b = \dfrac{h_s + h_p}{2}, & \text{(d)} & I = 2\left[\dfrac{w_b h_p^3}{12} + w_b h_p b^2 \right] + \dfrac{E_s / E_p w_b h_s^3}{12}, & \text{(e)} \end{array} \tag{4}$$

where k represents the effective elastic stiffness, k_1 and k_2 are the average strain to mass displacement and the input force to average induced stress, b is a geometrical parameter of the bimorph structure and I is the composite inertial moment of the beam. Usually, the internal resistance R_i of a piezoelectric crystal is very high, hence ω_i is negligible.

3.2. Case 2: electromagnetic generator

Let consider a simple "magnet in-line coil" electromagnetic generator as schematized in Figure 3a. we can rewrite the characteristic parameters as

$$\begin{array}{llll} \alpha = Bl / R_L, & \text{(a)} & \lambda = Bl = \alpha R_L, & \text{(b)} \\ \omega_c = R_L / L_e, & \text{(c)} & \omega_i = R_i / L_e, & \text{(d)} \end{array} \tag{5}$$

with B representing the magnetic field across the coil of total length l and self-inductance $L_{\rm e}$. Even in this case, by assuming an internal resistance of the coil R_0 small with respect to the external load ($R_0 << R_{\rm L}$), we can neglect ω_i. Actually, the above fundamental parameters are also valid for other systems such as "magnet across coil" arrangement depicted in Figure 3b.

In both cases the governing equations are the same, and they can be also rewritten in a more convenient nondimensional form similar to [22].

3.3. Transfer functions

Let now consider the simple case of harmonic excitation $\ddot{y}(t) = Y_0 e^{j\omega t}$ as input. We can transform the motion equations(1) and (2) into the Laplace domain, with $s=j\omega$ as the Laplace variable (where j stands for the imaginary unit). The function Y(s), Z(s), and V(s) are the acceleration amplitude, mass displacement and output voltage delivered to the resistive load respectively. Thus, the governing equations for the single-mass generator can be rewritten as the system

$$\begin{pmatrix} ms^2+ds+k & \alpha \\ -\lambda\omega_c s & s+\omega_c \end{pmatrix}\begin{pmatrix} Z \\ V \end{pmatrix}=\begin{pmatrix} -mY \\ 0 \end{pmatrix} \tag{6}$$

the left-side matrix, that we can name *A*, represents the generalized impedance of the oscillating system. By means of linear algebraic methods we can easily solve the above system equation, so that the displacement *Z*(s) and output voltage *V*(s) are given by

$$Z=\frac{-mY}{\det A}(s+\omega_c)=\frac{-mY\cdot(s+\omega_c)}{ms^3+(m\omega_c+d)s^2+(k+\alpha\lambda\omega_c+d\omega_c)s+k\omega_c}, \tag{7}$$

$$V=\frac{-mY}{\det A}\lambda\omega_c s=\frac{-mY\cdot\lambda\omega_c s}{ms^3+(m\omega_c+d)s^2+(k+\alpha\lambda\omega_c+d\omega_c)s+k\omega_c}. \tag{8}$$

Hence, the transfer functions between displacement and voltage over input acceleration are therefore given by

$$H_{ZY}(s)=\frac{Z}{Y}, \quad \text{(a)} \qquad\qquad H_{VY}(s)=\frac{V}{Y}. \quad \text{(b)} \tag{9}$$

By substituting $s=j\omega$ in (8), we can calculate the electrical power dissipated across the resistive load

$$\begin{aligned} P_e(\omega)&=\frac{|V(j\omega)|^2}{2R_L}=\frac{|H_{VY}(j\omega)|^2|Y(j\omega)|^2}{2R_L}= \\ &=\frac{Y_0^2}{2R_L}\left|\frac{-m\lambda\omega_c j\omega}{(\omega_c+j\omega)(-m\omega^2+j\omega d+k)+\alpha\lambda\omega_c j\omega}\right|^2. \end{aligned} \tag{10}$$

By introducing the natural frequency of the undamped oscillation $\omega_n=\sqrt{k/m}$ and the normalized damping factor $\zeta=d/2m\omega_n$ into the above equation, it becomes

$$P_e(\omega)=\frac{Y_0^2}{2R_L}\left|\frac{-\lambda\omega_c j\omega}{(\omega_c+j\omega)(\omega_n^2-\omega^2+2\zeta\omega_n j\omega)+\alpha\lambda\omega_c j\omega/m}\right|^2 \tag{11}$$

In Figure 7 the graph of the normalized power function over the nondimensional variable ω/ω_n of a generic vibration-based generator is shown.

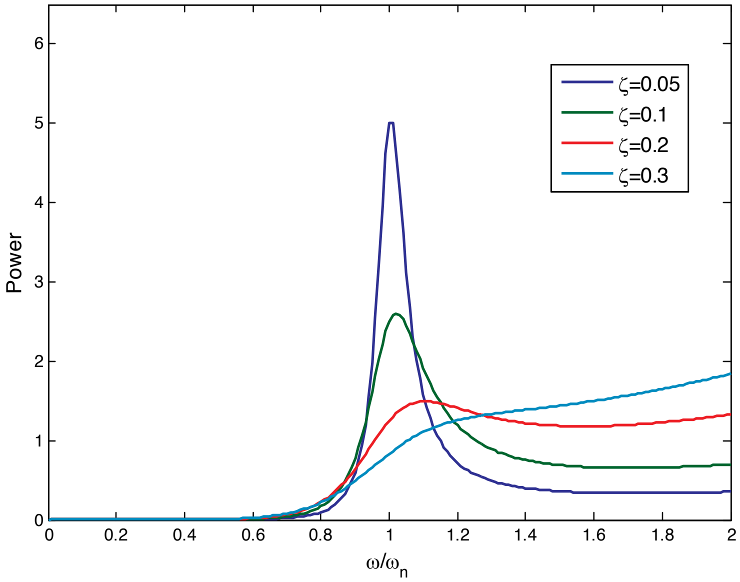


Figure 7. Normalized power function of a generic vibration-base harvester.

The main limits of a linear vibration energy harvester include:

- narrow bandwidth, that also implies applications with vibrating sources tuned around resonant frequency of the harvester,
- versatility and adaptation to variable spectrum vibration sources,
- at MEMS scale: small inertial mass and maximum displacement, limited power not suitable for milliwatts electronics (10-100mW).

In any case, for applications and environments that feature a vibration source consistent in frequency and time, linear systems can still represent an optimal choice.

4. Beyond linear energy harvesting

In most of the reported studies, the energy harvesters are designed as linear oscillators that match their resonant frequencies to the excitation frequencies of the environment to achieve the maximum output power.

This condition can be easily performed when the excitation frequency is well known and stable in time. It is in fact possible to choose the correct geometry and harvester dimension for frequency matching. However, when the ambient excitation frequency is unknown or varies in time, the previous conditions are not guaranteed. Therefore a harvester with a fixed resonant frequency is not able to achieve an optimal output power.

Various strategies have been investigated to overcome this practical inconvenient and increase the bandwidth of vibration-based harvesters. In the following the state-of-the-art techniques are summarised in three main categories: resonance frequency tuning, multimodal oscillators and frequency up-conversion. A complete technique review is presented by [23].

4.1. Frequency tuning

In some cases it has been demonstrated [24] that the resonance frequency of an oscillator can be tuned to the main exciting frequency in two different ways: passive and active. The passive mode requires an intermittent power input (manual or automatic) to check and tune the system until the frequency match is complete, then the power requirement is zero since the excitation frequency varies again. The active mode is more power demanding since a continuous power is needed to tune the system; this higher power consumption brings the effect to increase the harvester efficiency.

The tuning mechanisms can be realised mechanically using springs or screws, with magnets or using a piezoelectric material.

Few works have been presented in the last years showing possible manual parameter adjustments to change the harvester stiffness or its mass configuration. The oscillator stiffness is changed with a pre-loaded or a pre-deflected, performing a softening or a hardening of the system.

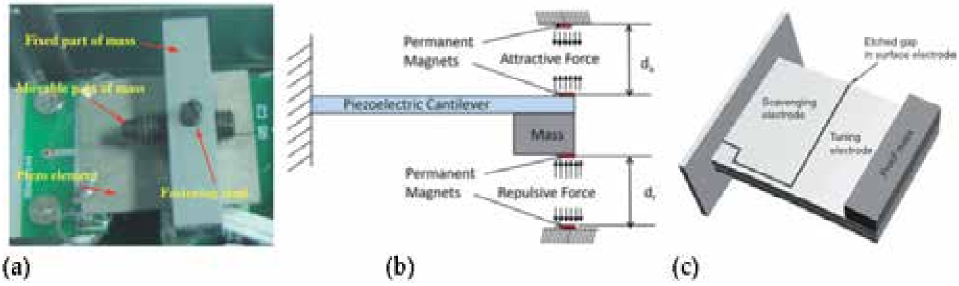


Figure 8. Three tunable vibration-based solutions: (a) Piezoelectric cantilever with a movable mass (source: Wu et al. 2008 [25]), (b) Piezoelectric cantilever with magnetic tuning (source: Challa et al. 2008 [26]), (c) Piezoelectric beam with a scavenging and a tuning part (source: Roundy and Zhang 2005 [27]).

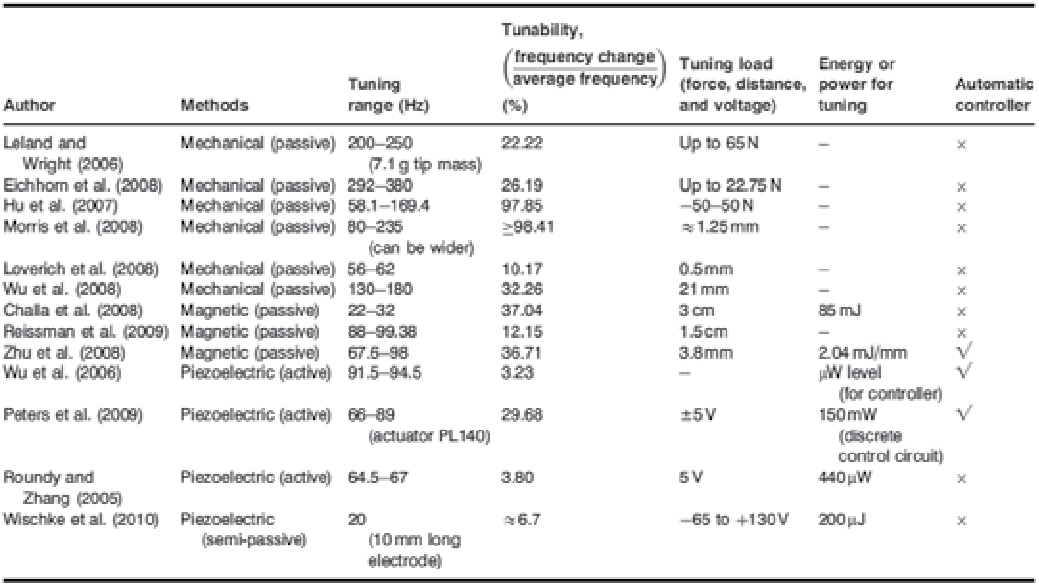
In Figure 8 three tunable solutions are presented. The three harvesters are realized with a piezoelectric beam, in which the tuning mechanism is purely mechanical in the first case by [25], magnetic in the second by [26], and piezoelectric in the third by [27].

The solution (a) is realized with a proof mass consisting in a fixed part combined with a movable part. The gravity centre of the proof mass can be adjusted by driving the screw. The fixed part of the mass is made with a relatively small density material and the tuning mass with a higher density one. In this way the frequency tunability is increased by moving the distance of the centre of gravity of the proof mass. In this prototype, the frequency range can be manually varied in the 130-180 Hz range moving the tip mass up to 21 mm.

The solution (b) of Figure 8 is realized by coupling two magnets fixed to the free end of the cantilever beam to two other magnets fixed to the top and bottom sides of the enclosure device. All the magnets have a vertical magnetization in such a way to perform an attractive and repulsive force on each side of the beam. Manually tuning the distance between the magnets using a screw, the magnetic force can be changed inducing a change in the cantilever stiffness. The resonant frequency can be varied in the 22-32 Hz range.

The latter case (c), of Figure 8, presented by Roundy and Zhang [25], is an active tuning mechanism. The electrode was etched to create both scavenging and tuning parts on the same beam. They analytically demonstrated that an active tuning actuator never resulted in a net increase in power output. This is explained because the power required to continuously tune the beam resonant frequency exceeds the harvested power increase.

Table 1, taken from Tang et al.[25], summarizes various tuning methods.



Author	Methods	Tuning range (Hz)	Tunability, $\left(\frac{\text{frequency change}}{\text{average frequency}}\right)$ (%)	Tuning load (force, distance, and voltage)	Energy or power for tuning	Automatic controller
Leland and Wright (2006)	Mechanical (passive)	200–250 (7.1 g tip mass)	22.22	Up to 65 N	–	×
Eichhorn et al. (2008)	Mechanical (passive)	292–380	26.19	Up to 22.75 N	–	×
Hu et al. (2007)	Mechanical (passive)	58.1–169.4	97.85	−50–50 N	–	×
Morris et al. (2008)	Mechanical (passive)	80–235 (can be wider)	≥98.41	≈1.25 mm	–	×
Loverich et al. (2008)	Mechanical (passive)	56–62	10.17	0.5 mm	–	×
Wu et al. (2008)	Mechanical (passive)	130–180	32.26	21 mm	–	×
Challa et al. (2008)	Magnetic (passive)	22–32	37.04	3 cm	85 mJ	×
Reissman et al. (2009)	Magnetic (passive)	88–99.38	12.15	1.5 cm	–	×
Zhu et al. (2008)	Magnetic (passive)	67.6–98	36.71	3.8 mm	2.04 mJ/mm	√
Wu et al. (2006)	Piezoelectric (active)	91.5–94.5	3.23	–	μW level (for controller)	√
Peters et al. (2009)	Piezoelectric (active)	66–89 (actuator PL140)	29.68	±5 V	150 mW (discrete control circuit)	√
Roundy and Zhang (2005)	Piezoelectric (active)	64.5–67	3.80	5 V	440 μW	×
Wischke et al. (2010)	Piezoelectric (semi-passive)	20 (10 mm long electrode)	≈6.7	−65 to +130 V	200 μJ	×

Table 1. Summary of various resonance tuning methods (source: Tang et al. 2010 [24]).

4.2. Multimodal energy harvesting

The multimodal approach consists in a multi-vibration harvester, designed to be excited when the natural driving frequency approaches one natural frequency of the harvester. In this case useful power can be harvested over multiple frequency spectra, increasing the bandwidth that can be covered for efficient energy harvesting.

One way to design a multimodal VEH consists in the combination of more transduction mechanisms together. In [24] a hybrid scenario was presented by Tadesse et al. as shown in Figure 9a. The harvester consists of a cantilever beam with piezoelectric crystal plates bonded on it at a fixed distance each other; a permanent magnet is attached at the cantilever tip oscillating within a coil fixed to the housing structure. It this configuration the electromagnetic transducer generates high output power at the cantilever first mode (at 20 Hz), while the piezoelectric transducer generates higher power at the cantilever second mode (at 300 Hz). The combination of the two schemes in one device is able to improve significantly the harvester response covering two frequency ranges. The drawback of this solution is the difficulty in combining the output power from two different mechanisms, thus requiring two separate converting circuits.

A different approach rather than exploiting the energy present at different modes of a single oscillator is to design a cantilever array integrated in one single device. If the geometric parameters of the harvester are appropriately selected, a wide vibration bandwidth can be exploited. In Figure 9b and 9c two different array solutions are presented.

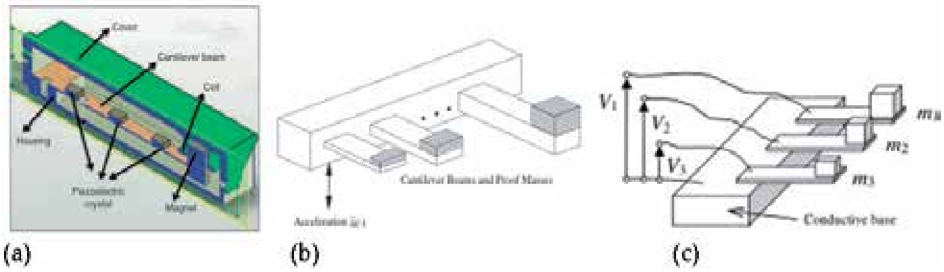


Figure 9. Multimodal VEH schematics. (a) Hybrid harvester with piezoelectric and electromagnetic transduction mechanisms (source: Tadesse et al. 2009 [28]). Piezoelectric cantilever arrays with various lengths and tip masses (b) (source: Shahruz 2006 [28]) and same cantilevers with different tip masses (c) (source: Ferrari et al. 2008 [29]).

In the first design, (b), various cantilevers with different lengths and tip masses attached to a common base compose a piezoelectric cantilever array. In the second, (c), the cantilevers are the same but the first mode response is changed varying the tip masses. Each cantilever presents a unique resonant frequency, the combination of which into a single device creates a sort of mechanical band-pass filter. By properly selecting the different parameters, the device can be designed to provide a voltage response on a wide frequency range.

It has been demonstrated by [30] that the improved bandwidth and performance were worth the modest increase in size of the proposed array device. A cantilever array configuration in respect to the hybrid solution, doesn’t present the difficulty in combining the output power from the different mechanisms, but requires one rectifier for each cantilever to avoid output cancellation due to the phase difference between the cantilevers.

As a matter of fact, the multimodal approach increases the bandwidth increasing the volume or the weight of the harvester, thus reducing the energy density. Specifically for the cantilever array, only one cantilever or a subset of them are active at the same time generating a certain amount of power while the others are at off-resonance. Hence, knowing the dominant spectrum of the ambient vibrations, the harvester has to be carefully designed to prevent a dramatic efficiency loss.

4.3. Up-conversion energy harvesting

In many practical situations, the ambient vibrations suitable for energy harvesting including human vibrations, natural events (i.e. wind, seismic motion), common household goods (i.e. fans, fridges, washing machine), automobiles, airplanes, structures etc. present their frequency content below few hundreds hertz.

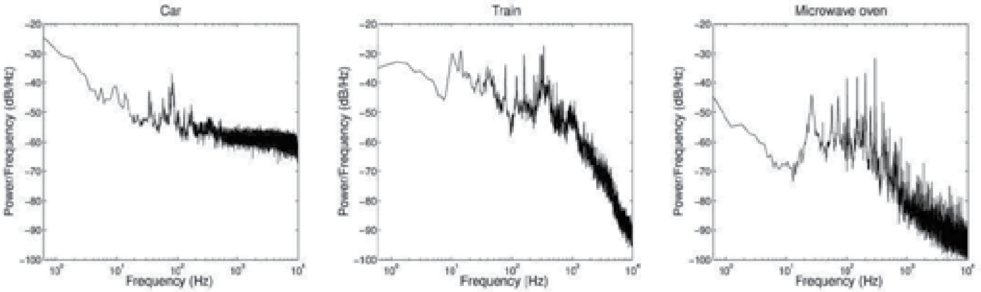


Figure 10. Vibration power spectra. Figure shows acceleration magnitudes (in db/Hz) vs frequency for three different environments.

As an example in Figure 10, three different spectra computed from vibrations taken from a car hood in motion, an operating microwave oven and a running train floor are presented.

All these different sources produce vibrations that are very large in amplitude and spectral characteristics. However, in these cases, like in great part of the cited examples, the ambient vibrations have their energy distributed over a wide frequency spectrum, with significant predominance of low frequency components and frequency tuning or a multimodal approach is not always possible due to geometrical/dynamical constraints. Hence, another frequency-robust solution for VEH is to 'transfer' the source vibration frequency to the harvester resonance frequency so that useful power can be harnessed in low frequency excitation scenarios.

A typical up-conversion schematic is presented in Figure 11, where the basic principle of such a device is evident [30]. The oscillator with elastic constant k has a resonant frequency in a lower region respect to the resonant frequencies of the piezoelectric beams. When the tooth passes and hits the cantilever tips, the cantilevers start oscillating at their natural frequency. Thus the low frequency vibration of the primary vibrating unit (i.e. the oscillating mass m) is transferred to the high frequency vibration of the secondary units (i.e. the piezoelectric cantilevers). This provides a robust low frequency harvesting using high frequency structures as transduction elements.

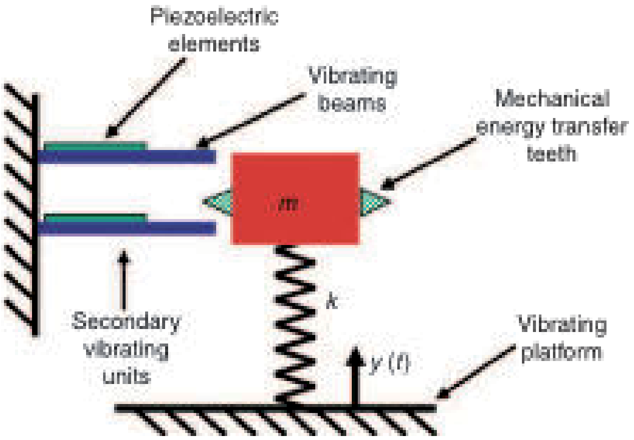


Figure 11. Schematic of an up-conversion VEH (source: Rastegar et al. 2006 [31]).

This frequency up-conversion technique was further pursued in few applications like the generators for low and variable speed rotary machineries [31]. It can be implemented combining different oscillators and interaction mechanisms, like magnetic or electrostatic, maintaining its basic principle.

This technique is a way to decouple the exciting frequency and the harvester vibration one, thus the performances are insensitive to the excitation frequency as long as it is less than the resonant frequency of the harvester. The drawback is that the ‘first unit’ resonance frequency needs to be tuned to the main frequency of the vibration source. In case the exciting vibrations are spread in a wide frequency range, the advantages of this method lose efficiency.

5. Non-linear techniques

All the above-mentioned strategies to harvest energy from the environment belong to the entire category of linear (or resonant) harvesting techniques. In general, for a generic linear system, even more complicated than a simple cantilever, the transfer function presents one or more peeks corresponding to the resonance frequencies and thus it is effective mainly when the incoming energy is abundant in that frequency region.

Unfortunately this is a serious limitation when it is required to build a vibration energy harvester of small dimension, for at least two main reasons, the first is that, as discussed above, the frequency spectrum of available vibrations instead of being sharply peaked at some frequency is usually very broad. The second reason is that the frequency spectrum of available vibrations is particularly rich in energy in the low frequency range, and it is very difficult, if not impossible, to build small, low-frequency resonant systems.

Based on these considerations it is clear that vibration harvesters inspired by cantilever-like configurations present a number of drawbacks that limit seriously their field of application.

The optimal vibration harvester characteristics for a broadband response can be summarized in the following points:

- Resonant oscillators should be avoided. In fact a resonant oscillator is capable of harvesting energy only in a very narrow band, around its resonant frequency. Non-resonant oscillators have to be taken into account.
- Increase the capability of harvesting energy at low frequency (below few hundred Hz) because this is where most of the ambient energy is available. Due to geometrical constraints a small dimension linear harvester in not feasible.
- No need for frequency tuning after the initial set-up of the harvester. Frequency tuning is a feature of resonant systems, thus if we move to non-resonant systems this requirement will be automatically satisfied.

As we have seen before the search for the best solution in terms of non-resonant systems should start from the potential energy function *U(z)*. In fact *U(z)* plays the role of dynamical energy storage facility (before transduction) for our mechanical oscillator and thus it is here that we should focus our attention.

The best solution in terms of non-resonant systems should start from the potential energy function *U(z)*. In fact equation (1) can be rewritten as:

$$m\ddot{z} + d\dot{z} + \frac{dU(z)}{dz} + \alpha V = -m\ddot{y}, \tag{12}$$

where *U(z)* plays the role of dynamical energy storage facility (before transduction) for our mechanical oscillator and thus it is here that we should focus our attention.

To replace a linear oscillator with a non-linear one the condition is:

$$U(z) \neq \frac{1}{2}kz^2 \tag{13}$$

meaning oscillators whose potential energy is not quadratically dependent on the relevant displacement variable. In recent years few possible candidates have been explored [32-42] running from:

$$U(z) = az^{2n} \tag{14}$$

to other more complicated expressions.

For non-linear oscillators it is not possible to define a transfer function, like in paragraph (3.3), and thus a properly defined resonant frequency even if the power spectral density can show one or more well defined peeks for specific values of the frequencies.

In this section, two of these non-linear potential cases, will be briefly addressed.

5.1. Bistable cantilever

An interesting option for a nonlinear oscillator is to look for a potential energy that is multi-stable, instead of mono-stable (like the linear case, i.e. the harmonic potential). A particularly simple and instructive example on how to move from the linear (mono-stable) to a possible nonlinear (bi-stable) dynamics is represented by a slightly modified version of the vibration harvester cantilever analysed above (see Figure 12).

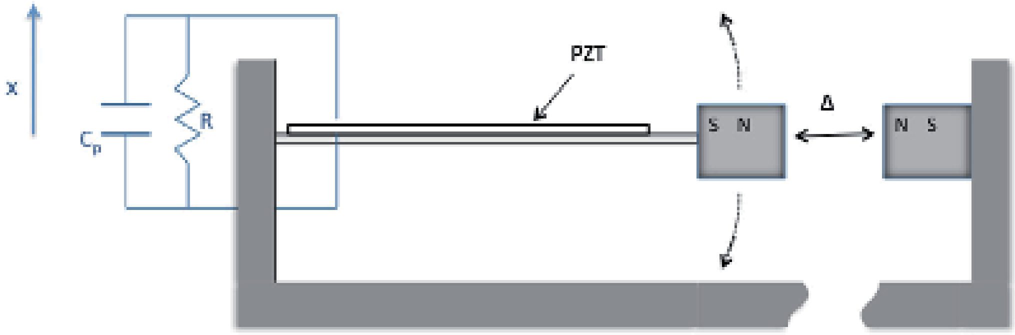


Figure 12. Schematic of the bi-stable energy harvester considered (source: Vocca et al., 2012 [32]).

This is a common cantilever with a bending piezoelectric layer. At the cantilever tip a small magnet is added. Under the action of the vibrating ground the pendulum oscillates alternatively bending the piezoelectric layer and thus generating a measurable voltage signal V. The dynamics of the inverted pendulum tip is controlled with the introduction of an external magnet conveniently placed at a certain distance Δ and with polarities opposed to those of the tip magnet. The interaction between the two magnets generates a force dependent from Δ that opposes the elastic restoring force of the bended beam. As a result, the inverted pendulum dynamics can show three different types of behaviours as a function of the distance Δ:

- when the two magnets are separated by a large distance ($\Delta >> \Delta_0$) the inverted pendulum behaves like a linear oscillator. This situation accounts well for the usual operating condition analysed previously.

- If Δ is small ($\Delta << \Delta_0$) the pendulum is forced to oscillate at the left or at the right of the vertical. In the limit of small oscillations, this can still be described in terms of a linear oscillator but with a resonant frequency higher that in the previous case.

- In between the two previous cases it exists an intermediate condition (Δ=Δ_0) where the cantilever swings in a more complex way with small oscillations around each of the two equilibrium positions (left and right of the vertical) and large excursions from one to the other.

The nonlinear potential that can be considered in equation (12) is [32]:

$$U(z)=\frac{1}{2}k_e z^2+(Az^2+B\Delta^2)^{-\frac{3}{2}} \tag{15}$$

with k_e, A and B representing constants related to the physical parameters of the cantilever. Clearly when the distance Δ between the magnets grows very large the second term in (15) becomes negligible and the potential tends to the harmonic potential of the linear case, typical of the cantilever harvester.

In Figure 13 the potential *U(z)* for the condition Δ=Δ_0 is shown. In this case the potential energy shows clearly two distinct equilibrium points separated by an energy barrier.

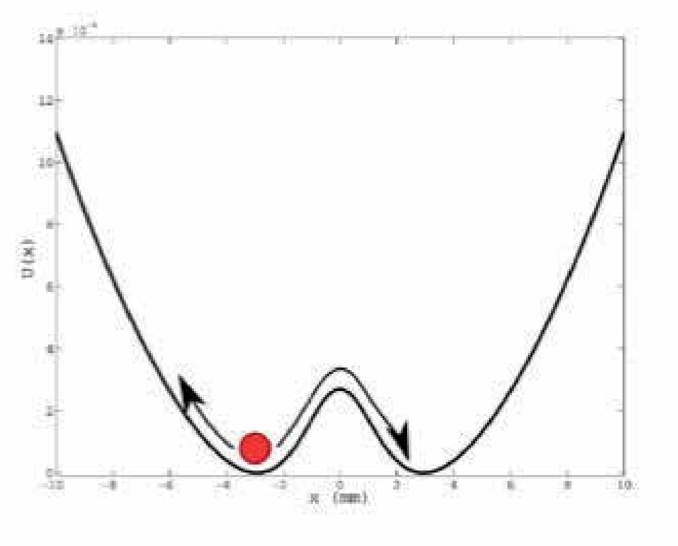


Figure 13. Potential energy *U(z)* in (0.15) in arbitrary units when (Δ=Δ_0).

The overall qualitative behaviour is somehow summarized in Figure 14, where the average power (P_{avg} = $V_{rms}{}^2/R$) extracted from this vibration harvester is presented as a function of the distance parameter Δ. As it is well evident there is an optimal distance Δ_0) where the power peaks to a maximum [33]. Most importantly such a maximum condition is reached in a full nonlinear regime (bistable condition of the potential) resulting larger (at least a factor 4) than the value in the linear condition (far right in Figure 14).

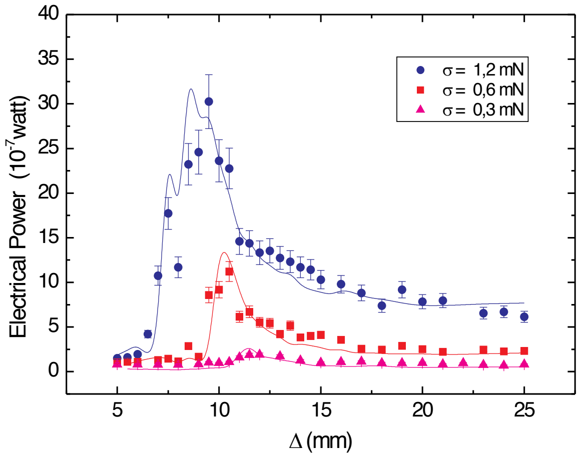


Figure 14. Piezoelectric nonlinear vibration harvester mean electrical power (P_{avg}=$V_{rms}{}^2/R$) as a function of the distance Δ between the two magnets (source: Cottone et al. 2009 [33]]).

5.2. Buckled beam

Buckled beams represent other possible structures to implement nonlinear VEH. In particular considering a piezoelectric beam clamped on both ends on a base excited vertically, it has demonstrated by Cottone et al.[33], that if the beam is subjected to a compression along his longitudinal axis (see Figure 15) the harvested power increases if the beam is vibrationally excited by a random noise.

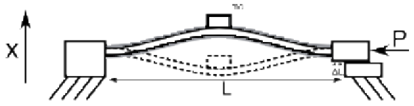


Figure 15. Schematic of a piezoelectric buckled beam (source: Vocca et al. 2013 [34]).

The equations that link the beam motion to the output voltage across a resistive load R_L are the followings [35]:

$$m\ddot{x} + \gamma\dot{x} + k_3x^3 + (k_2 - k_1V)x + k_0V = \sigma\xi, \tag{16}$$

$$\frac{1}{2}C\dot{V} + \frac{V}{R_L} = -k_1x\dot{x} + k_0\dot{x} \tag{17}$$

where m, γ, C and R_L are the equivalent mass, the viscous parameter, the coupling capacitance and the resistive load respectively. The term k_3 is the beam third order stiffness coefficient; k_0 and k_1 are the piezoelectric coupling terms due to the bending and to the axial strain of the beam respectively. The term $k_2 = k_a - k_b\Delta L$ is the stiffness parameter, where k_a and k_b are constants depending on physical parameters of the beam. When the beam is compressed by increasing ΔL (or equivalently the lateral load P), the stiffness becomes negative and the system becomes bistable. The buckled model in equation (16) is valid for small compressions ΔL. The $\sigma\xi(t)$ represents the vibration force that drives the beam.

It has been demonstrated in [34], that considering as excitation a random force with a Gaussian distribution, zero mean and exponentially auto-correlated, the output power that represents the amount of the energy harvested per second, increases increasing ΔL. In Figure 16 the average electrical power versus the compression length ratio respectively for the experimental (on the left) and numerical models (on the right) are shown for three noise amplitudes.

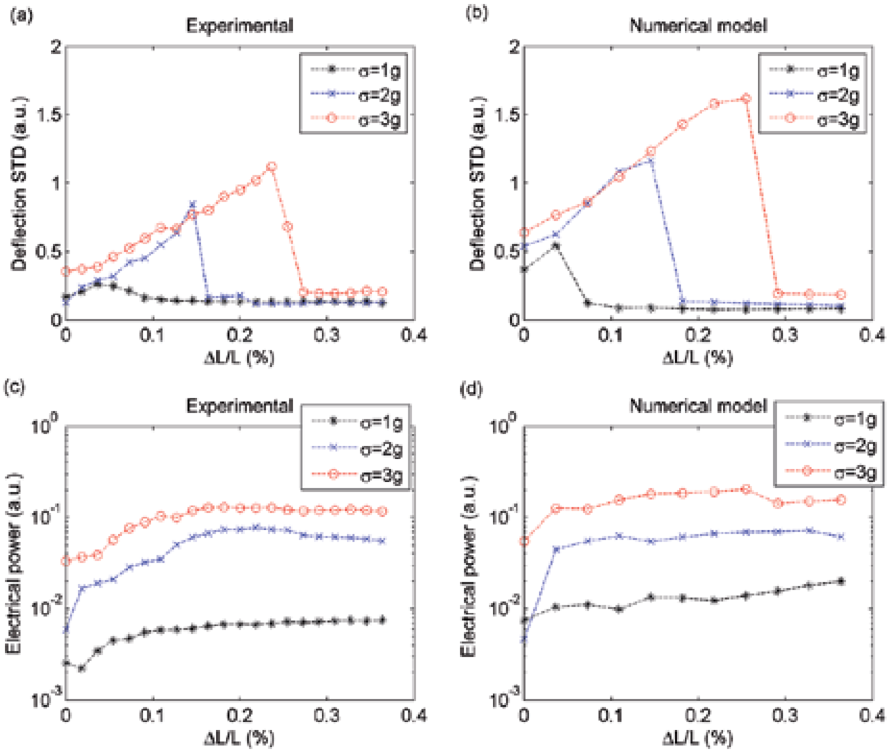


Figure 16. Electrical average power versus relative compression *ΔL/L (%)* for experiment (left column) and numerical simulation (right column). (Source: Cottone et al. 2012, [34]).

Moreover, the piezoelectric beam produces up to an order of magnitude more electric power when it is compressed than in the unbuckled case.

5.3. Comparison of the two methods

In Vocca et al. [34] as a comparison between the cantilever nonlinear configuration and the buckled beam, the same piezoelectric element subjected to a fixed vibrating body in both configurations has been simulated. The piezoelectric oscillators output responses obtained as a function of the nonlinear parameter are compared in Figure 17.

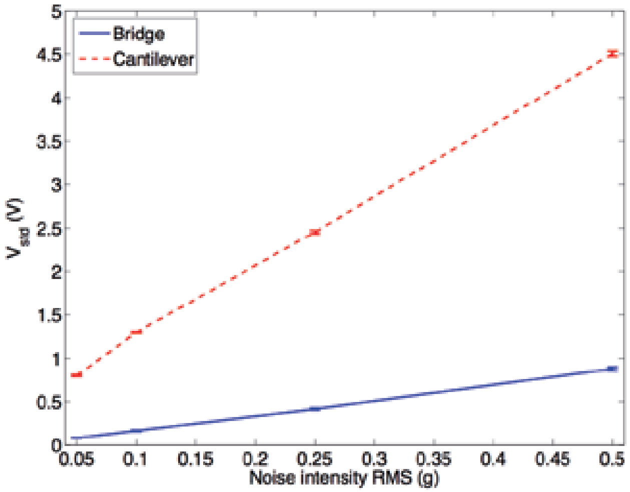


Figure 17. Comparison between cantilever and buckled beam response as a function of the noise intensity (source: Vocca et al. 2013, [35]).

The cantilever configuration appears to perform better than the beam in all the conditions. Although the buckled beam configuration is an interesting option for an harvester configuration where the introduction of a magnetic field is undesirable, generally it becomes evident that the cantilever harvester configuration is the best choice once in presence of an exponentially correlated noise.

6. Conclusions

In this chapter we have discussed the various strategies developed for vibration energy harvesting. A specific reference to the role of linearity and nonlinearity has been discussed. We have shown that linear resonant systems are clearly limited in their practical applications even if various techniques have been developed to increase the bandwidth. It has been shown that more complex VEHs based on non-linear mechanical oscillators can outperform them in a number of realistic energy harvesting scenarios. In fact, from comparative works that have been recently published it comes out that nonlinear mononstable and bistable structures are the best opinion for enhancing the overall performances and improving the flexibility of vibration powered electronics. [35]

Author details

Helios Vocca[1] and Francesco Cottone[2]

1 NiPS Laboratory, Department of Physics, University of Perugia, Perugia, INFN Perugia and Wisepower srl, Italy

2 ESIEE Paris, University of Paris Est, Paris, France

References

[1] C. S. Raghavendra, *et al.*, *Wireless sensor networks*: Springer, 2006.

[2] N. Gershenfeld, *et al.*, "The Internet of things," *Scientific American,* vol. 291, pp. 76-81, 2004.

[3] P. D. Mitcheson, *et al.*, "Energy harvesting from human and machine motion for wireless electronic devices," *Proceedings of the IEEE,* vol. 96, pp. 1457-1486, 2008.

[4] A. Bilbao, *et al.*, "Ultra-low power wireless sensing for long-term structural health monitoring," 2011, p. 798109.

[5] K. Cook-Chennault, *et al.*, "Powering MEMS portable devices—a review of non-regenerative and regenerative power supply systems with special emphasis on piezoelectric energy harvesting systems," *Smart materials and structures,* vol. 17, p. 043001, 2008.

[6] L. Wang and F. Yuan, "Energy harvesting by magnetostrictive material (MsM) for powering wireless sensors in SHM," 2007, pp. 18-22.

[7] T. Ikeda, *Fundamentals of Piezoelectricity*. Walton St, Oxford, UK: Oxford University Press 1996.

[8] A. b. E. Lefeuvre, C. Richard, L. Petit, D. Guyomar, "A comparison between several approaches of piezoelectric energy harvesting," *J. Phys. IV France,* vol. 128, pp. 177-186, 2005.

[9] T. G. McKay, *et al.*, "Soft generators using dielectric elastomers," *Applied Physics Letters,* vol. 98, pp. 142903-142903-3, 2011.

[10] G. Poulin, *et al.*, "Generation of electrical energy for portable devices Comparative study of an electromagnetic and a piezoelectric system," *Sensors & Actuators: A. Physical,* vol. 116, pp. 461-471, 2004.

[11] S. P. Beeby, *et al.*, "A micro electromagnetic generator for vibration energy harvesting," *Journal of Micromechanics and Microengineering,* vol. 17, p. 1257, 2007.

[12] P. Wang, *et al.*, "A micro electromagnetic low level vibration energy harvester based on MEMS technology," *Microsystem Technologies,* vol. 15, pp. 941-951, 2009.

[13] I. Sari, *et al.*, "An electromagnetic micro power generator for wideband environmental vibrations," *Sensors and Actuators A: Physical,* vol. 145, pp. 405-413, 2008.

[14] S. Meninger, *et al.*, "Vibration-to-electric energy conversion," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on,* vol. 9, pp. 64-76, 2001.

[15] F. Peano and T. Tambosso, "Design and optimization of a MEMS electret-based capacitive energy scavenger," *Microelectromechanical Systems, Journal of,* vol. 14, pp. 429-435, 2005.

[16] P. Basset, *et al.*, "A batch-fabricated and electret-free silicon electrostatic vibration energy harvester," *Journal of Micromechanics and Microengineering,* vol. 19, p. 115025, 2009.

[17] S. Roundy, *et al.*, "MICRO-ELECTROSTATIC VIBRATION-TO-ELECTRICITY CONVERTERS," *Fuel Cells (methanol),* vol. 220, p. 22.

[18] P. Miao, *et al.*, "MEMS inertial power generators for biomedical applications," *Microsystem Technologies,* vol. 12, pp. 1079-1083, 2006.

[19] X. Chen, *et al.*, "1.6 V nanogenerator for mechanical energy harvesting using PZT nanofibers," *Nano letters,* vol. 10, pp. 2133-2137, 2010.

[20] X. Wang, *et al.*, "Direct-Current Nanogenerator Driven by Ultrasonic Waves," *Science,* vol. 316, p. 102, 2007.

[21] C. B. Williams and R. B. Yates, "Analysis Of A Micro-electric Generator For Microsystems," *Solid-State Sensors and Actuators, 1995 and Eurosensors IX. Transducers' 95. The 8th International Conference on,* vol. 1, 1995.

[22] S. Roundy, *et al.*, *Energy Scavenging For Wireless Sensor Networks with special focus on Vibrations*: Kluwer Academic Publisher, 2004.

[23] G. Sebald, *et al.*, "Experimental Duffing oscillator for broadband piezoelectric energy harvesting," *Smart materials and structures,* vol. 20, p. 102001, 2011.

[24] L. Tang, *et al.*, "Toward broadband vibration-based energy harvesting," *Journal of Intelligent Material Systems and Structures,* vol. 21, pp. 1867-1897, 2010.

[25] S. Roundy and Y. Zhang, "Toward self-tuning adaptive vibration-based microgenerators," in *Proc. SPIE*, 2004, pp. 373-384.

[26] X. Wu, *et al.*, "A frequency adjustable vibration energy harvester," *Proceedings of PowerMEMS,* pp. 245-248, 2008.

[27] V. R. Challa, *et al.*, "A coupled piezoelectric–electromagnetic energy harvesting technique for achieving increased power output through damping matching," *Smart materials and structures,* vol. 18, p. 095029, 2009.

[28] Y. Tadesse, *et al.*, "Multimodal energy harvesting system: piezoelectric and electromagnetic," *Journal of Intelligent Material Systems and Structures,* vol. 20, pp. 625-632, 2009.

[29] S. Shahruz, "Design of mechanical band-pass filters for energy scavenging," *Journal of Sound and Vibration,* vol. 292, pp. 987-998, 2006.

[30] M. Ferrari, *et al.*, "Piezoelectric multifrequency energy converter for power harvesting in autonomous microsystems," *Sensors and Actuators A: Physical,* vol. 142, pp. 329-335, 2008.

[31] J. Rastegar, *et al.*, "Piezoelectric-based power sources for harvesting energy from platforms with low frequency vibration," in *Proc. SPIE*, 2006, p. 617101.

[32] H. Vocca, *et al.*, "Kinetic energy harvesting with bistable oscillators," *Applied Energy,* 2012.

[33] F. Cottone, *et al.*, "Nonlinear energy harvesting," *Physical Review Letters,* vol. 102, 2009.

[34] F. Cottone, *et al.*, "Piezoelectric buckled beams for random vibration energy harvesting," *Smart materials and structures,* vol. 21, 2012.

[35] H. Vocca, *et al.*, "A comparison between nonlinear cantilever and buckled beam for energy harvesting," *European Physical Journal,* 2013.

[36] L. Gammaitoni, *et al.*, "Nonlinear oscillators for vibration energy harvesting," *Applied Physics Letters,* vol. 94, pp. 164102-164102-3, 2009.

[37] L. Gammaitoni, *et al.*, "The benefits of noise and nonlinearity: Extracting energy from random vibrations," *Chemical Physics,* vol. 375, pp. 435-438, 2010.

[38] M. Ferrari, *et al.*, "Improved energy harvesting from wideband vibrations by nonlinear piezoelectric converters," *Sensors and Actuators A: Physical,* 2010.

[39] A. Arrieta, *et al.*, "A piezoelectric bistable plate for nonlinear broadband energy harvesting," *Applied Physics Letters,* vol. 97, p. 104102, 2010.

[40] B. Andò, *et al.*, "Nonlinear mechanism in MEMS devices for energy harvesting applications," *Journal of Micromechanics and Microengineering,* vol. 20, p. 125020, 2010.

[41] D. A. Barton, *et al.*, "Energy harvesting from vibrations with a nonlinear oscillator," *Journal of vibration and acoustics,* vol. 132, 2010.

[42] S. C. Stanton, *et al.*, "Nonlinear dynamics for broadband energy harvesting: Investigation of a bistable piezoelectric inertial generator," *Physica D: Nonlinear Phenomena,* vol. 239, pp. 640-653, 2010.

Chapter 4

Thermoelectric Energy Harvesting

Douglas Paul

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57092

1. Introduction

The generation of electrical energy from thermal energy was originally discovered by Thomas Johann Seebeck in 1822 when he first demonstrated that a thermoelectric voltage was produced after providing a temperature difference across two materials. Jean Charles Athanase Peltier then demonstrated in 1834 that the application of a current could be used to pump heat, an effect with great potential for refrigeration. It was not until the 1850s that Lord Kelvin worked out the physics of the Seebeck and Peltier effects attributing the reversible heat flow discovered by Peltier must have an entropy associated with it and the Seebeck coefficient was a measure of the entropy associated with the electric current. Further developments in the theoretical understanding of thermoelectrics required quantum mechanics. The efficiency of the thermoelectric generation process was derived in 1911 by Edmund Altenkirch.

2. Fundamental physics

Before describing and deriving the main thermodynamic properties and equations, it is worthwhile having a brief review of the key parts of physics required for thermoelectrics. One of the first effects that will be used to derive the thermoelectric efficiency is Joule's law of heating. Joule was the first to demonstrate that any current passing through a resistor produces an amount of heat (Fig. 1(a)). Specifically the heat, Q (as a power i.e. energy / time) generated by passing a current, I though a resistance, R is given by Joule's first law

$$Q = I^2R \tag{1}$$

It should be clear that thermoelectrics have heat being transported through a range of materials and some understanding of the transport of that heat is required. The heat generated by any process will be transported through a material driven by any temperature

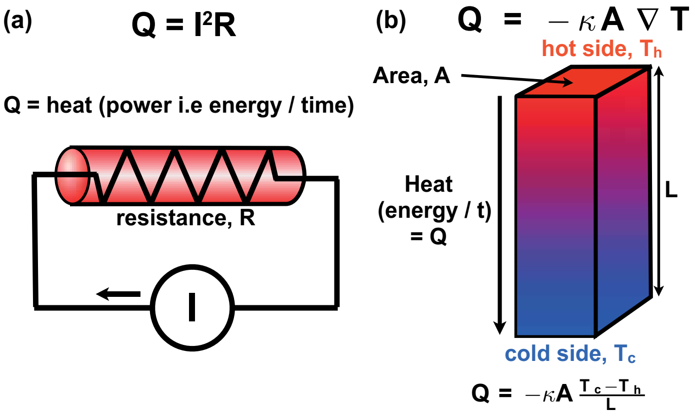


Figure 1. (a) The heat power generated by Joule heating of a resistance, R with current, I flowing through the resistance. (b) The thermal transport through a rod of area, A, length L and thermal conductivity κ as defined by Fourier's law of thermal transport.

gradients along the material (Fig. 1(b)). This is given by Fourier's law of heat transport which states for a material with area A and thermal conductivity κ that

$$Q = -\kappa A \nabla T \left(= -\kappa A \frac{T_c - T_h}{L} \text{for 1D transport along a length, } L \right) \tag{2}$$

Strictly speaking the heat transport could be in multiple directions in a complex material with a range of different thermal conductivities in different directions but for most thermoelectrics systems, the designs attempt to keep the heat flow simple by using 1 dimensional constructs. The right hand term of equation 2 is the equation for 1D heat transport which is predominantly the one used in thermoelectric devices and modules.

After discussing the generation and transport of heat, we can now discuss the thermoelectric effects. It is far easier to understand the Peltier effect and so this will be discussed before the Seebeck effect. We need to consider two materials in a thermocouple connected between a hot reservoir of temperature, T_h and a cold reservoir of temperature T_c (Fig. 2(a)). To produce the Peltier effect, a current has to be applied and so Fig. 2(a) demonstrate the system with the current being applied to the ends of material 2. The Peltier coefficient, Π, for this system is given by

$$\Pi = \frac{Q}{I} \tag{3}$$

The units of Π are the Volt since heat divided by current is Watts divided by Amps. The physics of what is going on is relatively simple. The Peltier coefficient is the heat energy carried by each electron per unit charge and time from the hot reservoir to the cold reservoir.

The Seebeck effect requires a similar circuit to be constructed but this time, the gap in material two is left open circuit (Fig. 2(b)). The open circuit voltage is proportional to

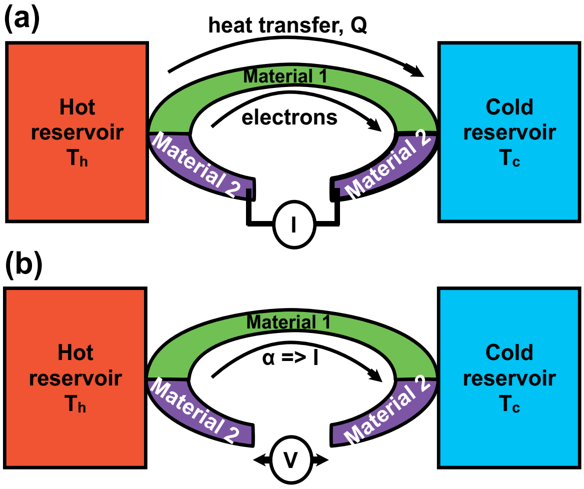


Figure 2. (a) The thermocouple system between two heat reservoirs required to demonstrate the Peltier effect (b) The thermocouple system between two heat reservoirs required to demonstrate the Seebeck effect.

$(T_h - T_c) = \Delta T$ and the constant of proportionality is called the Seebeck coefficient, α. More generally the Seebeck coefficient is defined as

$$\alpha = \frac{\mathrm{d}V}{\mathrm{d}T} \tag{4}$$

The units for the Seebeck coefficient are V/K. The Seebeck coefficient is 1/q times the entropy (Q/T) transported with each electron where q is the electron charge. Hence the Peltier effect is just each electron in the electrical current transferring an amount of heat from one reservoir to the other i.e. a heat pump.

To calculate the Seebeck or Peltier coefficients from theory requires one to solve the Boltzmann transport equation in the relaxation time approximation. This is beyond the scope of this text but a full derivation can be found in a range of solid state text books including [1]. From this approach, the Seebeck coefficient can be written in terms of the energy of an electron, E, Boltzmann's constant, k_B, the Fermi level in the thermoelectric material, E_F and the momentum relaxation time, τ, as

$$\alpha = -\frac{k_B}{q}\int \frac{(E - E_F)}{k_B T}\frac{\sigma(E)}{\sigma}\mathrm{d}E \tag{5}$$

where the electrical conductivity

$$\sigma = \int \sigma(E)dE = q\int g(E)\mu(E)f(E)(1 - f(E))\mathrm{d}E \tag{6}$$

The functions are the density of states, $g(E)$, the carrier mobility, $\mu(E)$ and the Fermi function, $f(E)$. The best thermoelectric materials are always semiconductors and so the equation for the Seebeck coefficient can be integrated over either the conduction band or the valence band to find the solution for n-type semiconductors and p-type semiconductors respectively. If we only consider electrons in the conduction band for energies above the conduction band edge, E_c then we have

$$\alpha = -\frac{k_B}{q}\left[\frac{(E_c - E_F)}{k_B T} + \frac{\int\limits_0^\infty \frac{(E-E_c)\sigma(E)\mathrm{d}E}{k_B T}}{\int\limits_0^\infty \sigma(E)\mathrm{d}E}\right] \quad \text{for } E > E_c \tag{7}$$

This equation is still quite complex and it is difficult to see exactly how to optimise thermoelectric materials. Cutler and Mott [2] realised a more useful form for the Seebeck coefficient. By differentiating the Fermi function it is easy to show that $f(1-f) = -k_B T\frac{\mathrm{d}f}{\mathrm{d}E}$ and then expanding $g(E)\mu(E)$ in a Taylor's series around $E = E_F$ it can be shown that

$$\alpha = -\frac{\pi^2}{3q}k_B^2 T\left[\frac{d\ln(\mu(E)g(E))}{dE}\right]\Bigg|_{E=E_F} \tag{8}$$

This equation is only valid for degenerately doped material, that is material that has doping above the Mott criteria and so the Fermi energy is greater than the conduction band edge. The doping density, n_c given by the Mott criteria is

$$n_c \approx \left(\frac{0.27}{a_B^*}\right)^3 \tag{9}$$

where a_B^* is the effective Bohr radius given by

$$a_B^* = \frac{\epsilon_0\epsilon_r h^2}{\pi m^* q^2} \tag{10}$$

with ϵ_0 the permittivity of free space (8.85×10^-12 Fm^{-1}), ϵ_r the dielectric constant of the semiconductor, h is Planck's constant and m^* is the effective mass of the charge carrier in the semiconductor. In metal-insulator theory this makes the degenerate semiconductor doped above Mott criteria metallic. The Cutler and Mott equation 8 now starts to suggests methods for optimising thermoelectric materials. Materials where the mobility and/or the density of states are varying by large amounts around the Fermi level have high Seebeck coefficients.

Further insights into how to increase the Seebeck coefficient can be found by taking the approach by Ziman [1]. If we ignore energy dependent scattering so that the momentum

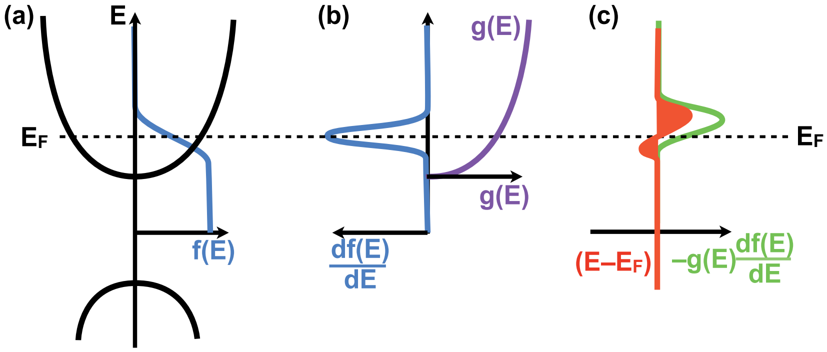


Figure 3. (a) The Fermi function demonstrating the carrier occupation as a function of energy for a degenerately doped semiconductor. (b) The first derivative of the Fermi function with respect to energy and also the density of states for the same semiconductor. (c) The product of the terms in (b) times $(E - E_F)$ which is related to the Seebeck coefficient as detailed in the text. It is the asymmetry between the two red areas above and below the Fermi energy that determines the magnitude of the Seebeck coefficient of the semiconductor.

relaxation time is given by $\tau(E)$ and the electron velocity by $v(E)$ then the electrical conductivity can be written as

$$\sigma = \frac{q^2}{3} \int \tau(E) v^2(E) \left[-g(E) \frac{\mathrm{d}f}{\mathrm{d}E} \right] \mathrm{d}E \tag{11}$$

Zimen then demonstrated that the Seebeck coefficient can be written as

$$\alpha = \frac{q^2}{3T\sigma} \int \tau(E) v^2(E) \left[-g(E) \frac{\mathrm{d}f}{\mathrm{d}E} \right] (E - E_F) \mathrm{d}E \tag{12}$$

Fig. 3 provides a graphical plot of the terms $\left[-g(E) \frac{\mathrm{d}f}{\mathrm{d}E} \right] (E - E_F)$ in this equation. With this approach by Ziman, σ, τ and v are all constant and it is therefore the asymmetry between the value of this term above and below the Fermi level that determines the magnitude of the Seebeck coefficient. This results is important when low dimensional structures are considered as large discontinuities of the density of states can potential provide significant enhancements to the Seebeck coefficient.

William Thomson (Lord Kelvin), realised that the Seebeck coefficient varies with temperature (Fig. 4) and so heat is both absorbed and generated in the thermocouples shown in Fig. 2. The gradient of the heat flux is then given by

$$\frac{dQ}{dx} = \beta I \frac{\mathrm{d}T}{\mathrm{d}x} \tag{13}$$

where β is the Thomson coefficient. Kelvin then derived the Kelvin relationships that hold for all materials which are

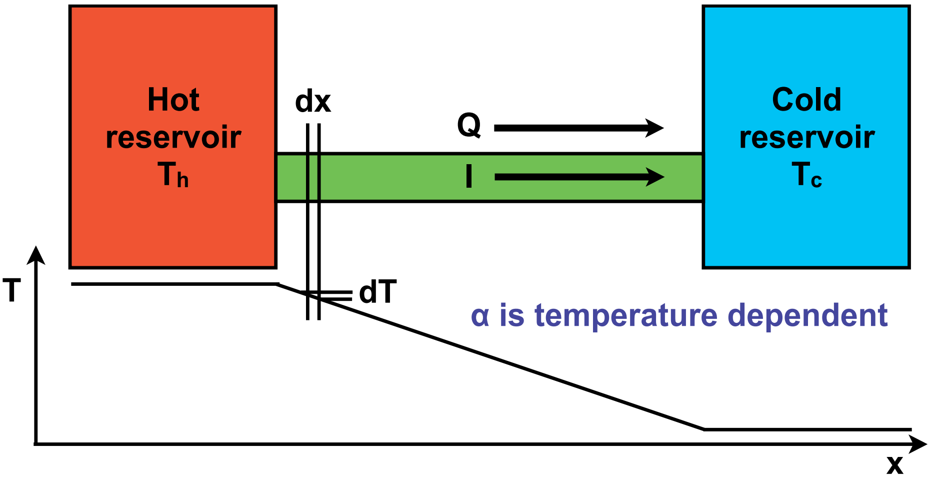


Figure 4. The Thomson coefficient is required as there will be a temperature dependence along any thermoelectric material connected to two heat reservoirs at different temperatures and this produces different Seebeck coefficients along the thermoelectric material. In this diagram the Seebeck coefficient varies along the x-direction i.e. $\alpha = \alpha(x)$

$$\Pi = \alpha T \tag{14}$$

$$\beta = T\frac{\mathrm{d}\alpha}{\mathrm{dT}} \tag{15}$$

The Peltier and Thomson coefficients are extremely difficult to measure experimentally but the Seebeck coefficient is relatively easy as it only requires a voltage measurement as a function of ΔT across the thermoelectric material. The Kelvin relationships therefore allow the measurement of the Seebeck effect to obtain the Peltier and Thomson coefficients.

3. Thermodynamic efficiency

We are about to derive the thermodynamical efficiency of thermoelectrics as first demonstrated by Altenkirch in 1911. Before undertaking this, it is sensible to review the Carnot cycle efficiency as this is the maximum efficiency for converting a given amount of thermal energy into work done. It will therefore set a maximum amount for any thermal system and will allow us to determine how much scope there may be for improving thermoelectric materials.

Fig. 5 shows the Carnot cycle where the input work done, W_{com} to a compressor increases the pressure of the water. This input work done is adiabatic so there is no gain or loss of heat within the complete system. The water flows from the compressor to a furnace where an amount of heat, Q_1 is input into the water at constant temperature (i.e. along an isotherm) so that the water is converted from water into dry steam. Therefore all the heat energy is being absorbed as the latent heat in this process of changing water into dry steam. The steam has a larger volume than the liquid water and so the volume increases as shown in Fig. 5(b). This

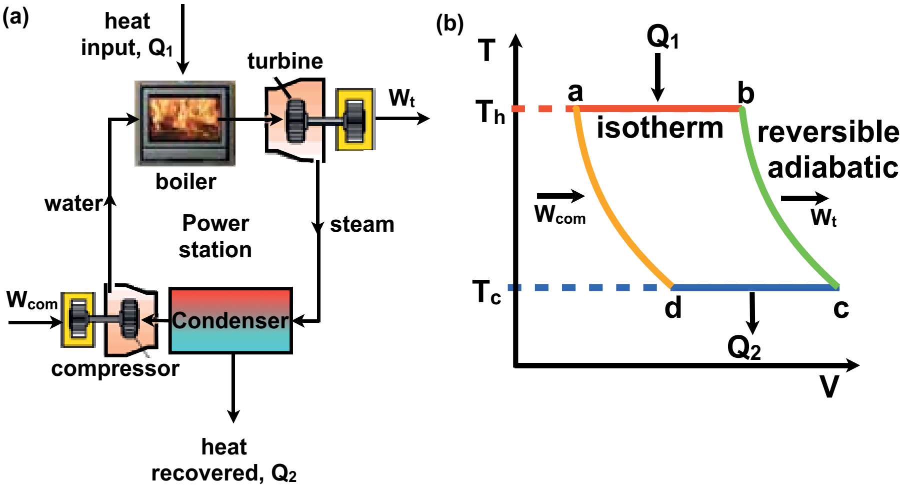


Figure 5. (a) The Carnot thermodynamical cycle showing a closed cycle water / steam system where input work, W_{com} is done of the water / steam by a compressor and the work done as output, W_t is that which is the output from a turbine. (b) The temperature-volume phase diagram for the Carnot thermodynamic cycle shown in (a).

increase in volume can be used to turn a turbine and the kinetic energy from the volume expansion can be recovered as work done on the turbine, W_t. The temperature is reduced in this process. Then to allow the cycle to start again, the steam has to be condensed into water, and the latent heat, Q_2 which is removed at constant temperature can be recovered and reused as an input. The process can then start again.

The efficiency of the Carnot cycle is given by

$$\text{Efficiency} = \eta = \frac{\text{net work output}}{\text{net work input}} = \frac{W_t - W_{com}}{Q_1} \tag{16}$$

From the first law of thermodynamics (conservation of energy), we have

$$(Q_1 - Q_2) - (W_t - W_{com}) = 0 \tag{17}$$

and so the efficiency can be rewritten as

$$\eta = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} \tag{18}$$

Carnot demonstrated using the temperature versus volume diagram for the Carnot cycle (Fig. 5(b)) that the maximum efficiency is only dependent on the maximum (T_h) and minimum temperatures (T_c) in the cycle and so the maximum efficiency known as the Carnot efficiency becomes

$$\eta_c = 1 - \frac{T_c}{T_h} \tag{19}$$

The Carnot efficiency is related to the Kelvin statement of the second law of thermodynamics which states that no system operating in a closed cycle can convert all the heat absorbed from a heat reservoir into the same amount of work. This is just a statement that no thermodynamic heat engine is 100 % efficient. The equivalent Caussius statement is that no process is possible whose sole result is the transfer of heat from a colder to a hotter body when no work is done inside the system. Heat must always flow from a hotter to a colder body. Heat can only be pumped from a colder to a hotter body by undertaking work on the system. Hence the Peltier effect provides one mechanism to pump heat when a current and electrical energy provides the work.

The equation clearly demonstrates that the efficiency can be increased by decreasing T_c or increasing T_h. More correctly the larger $\Delta T = T_h - T_c$ becomes, the higher the efficiency. Therefore for practical systems, the easiest way to increase the efficiency of any heat engine is to increase the hot reservoir temperature, T_h.

We have already described how a temperature gradient across a material results in heat conduction through Fourier's law when no electrical current flows in the system. If there is an electrical current in the same direction due to the Seebeck effect then the Peltier effect will attempt to oppose the applied temperature gradient. Therefore when the heat flows through a thermoelectric material between hot and cold reservoirs (see Fig. 1(a)), we have to consider not just the Fourier heat transport but also the Peltier effect.

We therefore need to write: Heat flux per unit area = Peltier term + Fourier term

$$\frac{Q}{A} = \Pi J - \kappa \nabla T \tag{20}$$

From the Kelvin relations we have $\Pi = \alpha T$ and the current density, $J = \frac{I}{A}$. Therefore this can be rewritten as

$$Q = \alpha I T - \kappa A \nabla T \tag{21}$$

We will now derive the thermodynamic efficiency of a thermoelectric generator. Before we can generate electricity, we need to build a circuit that can deliver power to a load. Fig. 6(a) shows the basic thermoelectric circuit for generating electricity which consists of an n-type and a p-type semiconductor connected electrically in series and thermally in parallel. The output power is delivered to a load which in this case will just be a resistor, R_L. For a Peltier cooler, a similar circuit can be designed where the load is replaced by a current source or battery as demonstrated in Fig. 6(b). The thermodynamic efficiency of the thermoelectric generator is given by

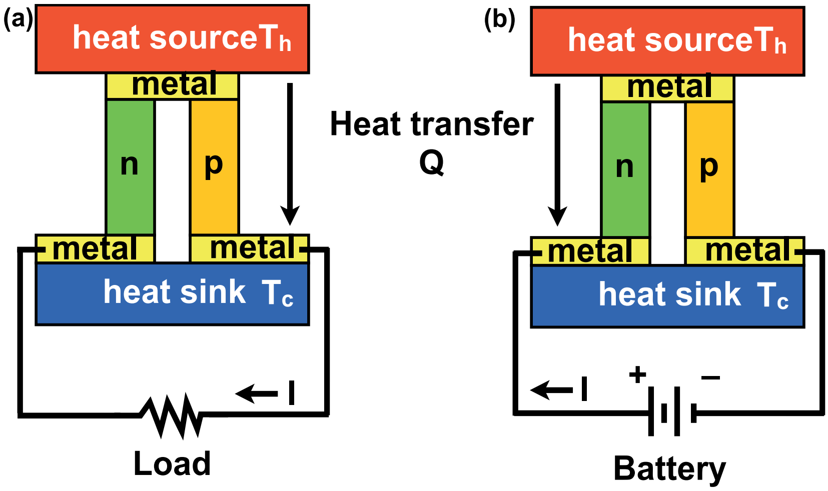


Figure 6. (a) A thermoelectric generator delivering power through current to a resistive load connected in electrical series to n- and p-type semiconductor thermoelectric materials. (b) A heat pump (or Peltier cooler) where a current is used to transport heat from the hot reservoir to the cold reservoir. By reversing the current, heat from the cold sink can be pumped by the Peltier effect to the hot sink by undertaking work.

$$\eta = \frac{\text{power supplied to load}}{\text{heat absorbed at hot junction}} \tag{22}$$

The power supplied to the load is just the Joule heating of the load resistor, R_L which is equal to I^2R_L. The heat absorbed at the hot junction is the Peltier term plus the heat withdrawn from the hot junction as described above. The Peltier heat is given by $\Pi I = \alpha I T_h$. If the resistance of the n-type and p-type semiconductor elements in series is R, then the current, I flowing in the circuit is just given by Ohm's Law as

$$I = \left(\frac{V}{R_{total}}\right) = \frac{\alpha(T_h - T_c)}{R + R_L} \tag{23}$$

The heat withdrawn from the hot junction is given by the Fourier term but as there will be Joule heating from the generated current from the Seebeck voltage, some heat will also be generated and returned to the hot junction. It is usual to assume that half of the Joule heat will be transport and half will be returned to the hot junction and so

$$\begin{aligned} Q_h &= \text{Fourier - Joule heating + half Joule heating returned} && (24)\\ &= \kappa A(T_h - T_c) - I^2R + \frac{1}{2}I^2R && (25)\\ &= \kappa A(T_h - T_c) - \frac{1}{2}I^2R && (26) \end{aligned}$$

We can now move to calculate the efficiency by combining these terms and assuming that the power supplied to the load is only through Joule heating to produce

$$\eta = \frac{\text{power supplied to load}}{\text{heat absorbed at hot junction}} \tag{27}$$

$$= \frac{\text{power supplied to load}}{\text{Peltier + heat withdrawn from hot junction } (Q_h)} \tag{28}$$

$$= \frac{I^2 R_L}{\alpha I T_h + \kappa A (T_h - T_c) - \frac{1}{2} I^2 R} \tag{29}$$

To find the maximum efficiency, this equation requires to be solved for $\frac{d\eta}{d\left(\frac{R_L}{R}\right)} = 0$. With a little algebra it can be shown that the solution is

$$\eta_{max} = (1 - \frac{T_c}{T_h}) \frac{\sqrt{1 + ZT} - 1}{\sqrt{1 + ZT} + \frac{T_c}{T_h}} \tag{30}$$

where $T = \frac{1}{2}(T_h + T_c)$ and the figure of merit for thermoelectrics is defined as

$$ZT = \frac{\alpha^2 \sigma}{\kappa} T \tag{31}$$

Equation 30 has two parts. The first part is just the Carnot efficiency given by $(1 - \frac{T_c}{T_h})$. The second part accounts for losses and irreversible processes which reduce as the dimensionless figure of merit, ZT, increases in value.

Fig. 7 demonstrates the Carnot efficiency and the maximum thermoelectric efficiencies for different ZTs as a function of ΔT. Also included are typically efficiencies for other thermodynamic cycles such as the Rankine and Stirling cycles for different thermal heating schemes. It is clear from this figure that thermoelectrics have significantly less efficiency today that Rankine or Stirling cycle engines. This is certainly true for power generation at the large scale. Below about 100 W, however, the Rankine and Stirling cycles become more difficult to sustain at high efficiencies and thermoelectrics have some advantages. This is mainly due to the fact that fluids effectively become more viscous (lossy) when dimensions are reduced below a certain length scale. A second major advantage of the thermoelectric generators is that they have no moving mechanical parts and are therefore significantly more reliable than Rankine or Stirling cycle engines which have compressors and turbines. Indeed, this is the major reason NASA used radioisotope thermoelectric generators for the Voyager space probes which have been operating for over 34 years and have now left the solar system.

So far we have made the assumption that the electrical and thermal properties of the n-type and p-type semiconductor legs are identical. This is seldom, if ever, the case in real thermoelectric materials. ZT needs to be redefined with the Seebeck coefficients, electrical

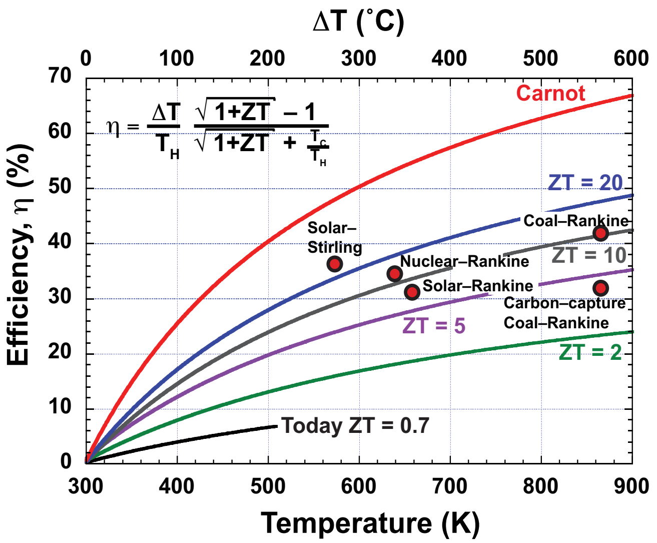


Figure 7. The thermodynamic efficiency of thermoelectric materials as a function of ZT assuming a cold side temperature of 298 K (25 ^{o}C). Also included is the Carnot efficiency and comparisons to Rankine and Stirling thermodynamic cycles.

conductivities and thermal conductivities defined for both n-type and p-type semiconductors to give

$$ZT = \frac{(\alpha_p - \alpha_n)^2 T}{\left[\sqrt{\frac{\kappa_p}{\sigma_p}} + \sqrt{\frac{\kappa_n}{\sigma_n}}\right]^2} \tag{32}$$

For n-type and p-type semiconductor legs of length (area), L_n (A_n) and L_p (A_p) respectively, ZT is a maximum value when the total resistance of the legs times the thermal conductance is a minimum value. This occurs when

$$\frac{L_n A_p}{L_p A_n} = \sqrt{\frac{\sigma_n \kappa_n}{\sigma_p \kappa_p}} \tag{33}$$

Up to this point we have assumed that we have a ΔT applied across each of the n- and p-type legs of the thermoelectric module. Since each leg has a given thermal conductivity, only a finite ΔT can be sustained across the thermoelectric legs. This therefore sets a maximum ΔT that can be sustained due to the thermal conductivity and in terms of ZT it is defined as

$$\Delta T_{max} = \frac{1}{2} Z T_c^2 \tag{34}$$

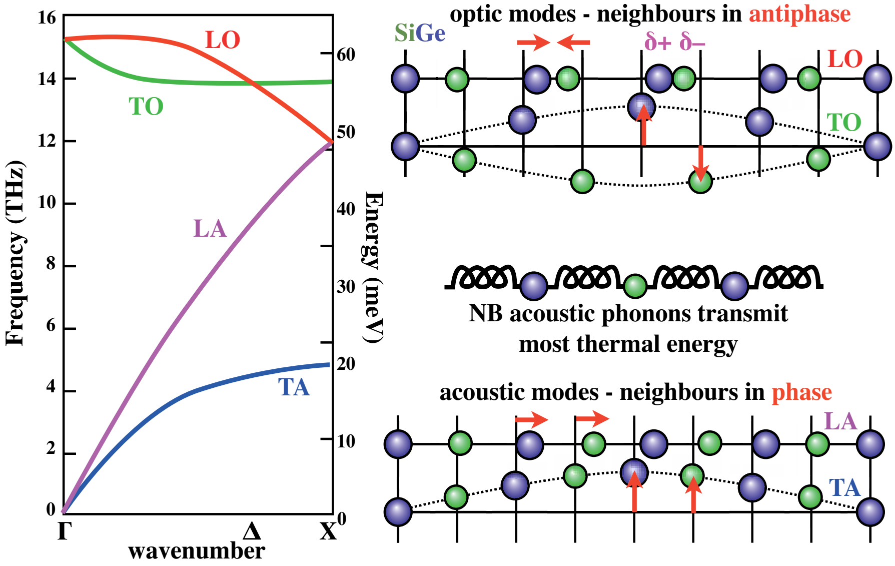


Figure 8. Left: The phonon energy dispersion versus wavenumber for Si. Right: schematic diagrams of the optic modes and acoustic modes in terms of the spring model for the lattice bonds between atoms in the crystal for an alloy of $Si_{0.5}Ge_{0.5}$.

4. Thermal conductivity

The thermal conductivity is one of the key parameters that many researchers aim to reduce to improve the ZT of a material. The thermal and electrical conductivities in bulk materials are linked as was first demonstrated by Wiedemann and Franz. They made the empirical observation that the thermal conductivity divided by the electrical conductivity times temperature was a constant for all metals. One of the greatest successes of the Drude theory of metals was the explanation of the Wiedemann-Franz law as it is now called. The Drude model assumes that the bulk of the thermal transport in metals is by the conduction electrons. This is now know to be incorrect and the success of the Drude model in explaining the Wiedemann-Franz law was a fortuitous cancelation of two factors of 100. It is now known that the Drude approach of applying the classical gas laws cannot be applied to electron gases in solids. The Wiedemann-Franz law, however, is still correct for degenerately doped bulk semiconductors and metals and provides substantial limitations when trying to optimize thermoelectric materials.

Quantum theory now provides a more complete description of the thermal conductivity which will be described. Phonons are the modes of vibrations of interacting particles in elastic crystal lattices. Phonons are quasi-particles which describe the collective excitations of the lattice (modes of vibrations). Fig. 8 shows the simple semi-classical interpretation where the bonds between atoms in a lattice are considered as springs and then by solving the Helmholtz wave equation for the vibrational modes of these springs throughout the lattice, the energy dispersion curves for the phonons can be calculated. Whilst more accurate description of the phonons requires much more detailed quantum mechanical calculations, this simple picture provides the same overall physical picture of the modes. There are two types of modes: optic and acoustic. The acoustic modes are where the oscillations of neighbours are in phase either in a transverse or longitudinal direction whilst the optics modes have the neighbours oscillating in anti-phase and are therefore at higher energy.

The total thermal conductivity of a semiconductor or metal can be divided into the electrical contribution, κ_{el} and the lattice contribution from phonons, κ_{ph}, i.e.

$$\kappa = \kappa_{el} + \kappa_{ph} \tag{35}$$

For non-degenerate semiconductors (low carrier density), $\kappa_{el} \ll \kappa_{ph}$ whilst for degenerate semiconductors (high carrier densities including metals) , $\kappa_{el} \gg \kappa_{ph}$. An ideal thermoelectric material should have both high carrier density and a decoupling of the thermal conductivity with $\kappa_{el} \ll \kappa_{ph}$ but Wiedemann-Franz prevents this in bulk semiconductors and metals.

In metals where κ_{el} dominates the the thermal conductivity, Wiedemann-Franz produces

$$\frac{\sigma T}{\kappa} = \frac{3}{\pi^2}\left(\frac{q}{k_B}\right)^2 = \frac{1}{L} \tag{36}$$

where L is the Lorenz number which is equal to 2.44×10^{-8} WΩK^{-2}. There are a number of examples where the Wiedemann-Franz law breaks down which include pure metals at low temperature, alloys where the small κ_{el} from high electron scattering results in significant contributions from κ_{ph} and certain low dimensional structures where κ_{ph} can dominate over κ_{el}.

The lattice contribution is in quantum theory the phonon contribution to the thermal conductivity where phonons are the quantised vibration of the lattice. The phenomenological model using phonon scattering which is now used to calculate the phonon contribution to the thermal conductivity this that first published by Callaway [3]. It assumes that the phonon scattering processes can be represented by frequency-dependent relaxation times and uses a formula derived from the Boltzmann transport equation to calculate the thermal conductivity. The lattice thermal conductivity calculated by this approach is

$$\kappa_{ph} = \frac{k_B}{2\pi^2}\left(\frac{k_B}{\hbar}\right)^3 \int_0^{\frac{\theta_D}{T}} \frac{\tau_c(x)x^4e^x}{v(x)(e^x-1)^2}dx \tag{37}$$

where θ_D is the Debye temperature, $x = \frac{\hbar\omega}{k_BT}$, $\hbar$ is Planck's constant divided by 2π, ω is the phonon angular frequency, τ_c is the combined phonon scattering time and v is the phonon velocity. The integral has to include all the acoustic and optical phonon modes although there are particular types of systems where for example transport of optical phonons can be forbidden in small enough nanostructures.

The electrical contribution to the thermal conductivity was derived from the Boltzmann transport equation by Nag [4]. For total electron momentum relaxation times of τ for electrons of energy, E, the electron contribution to the thermal conductivity is

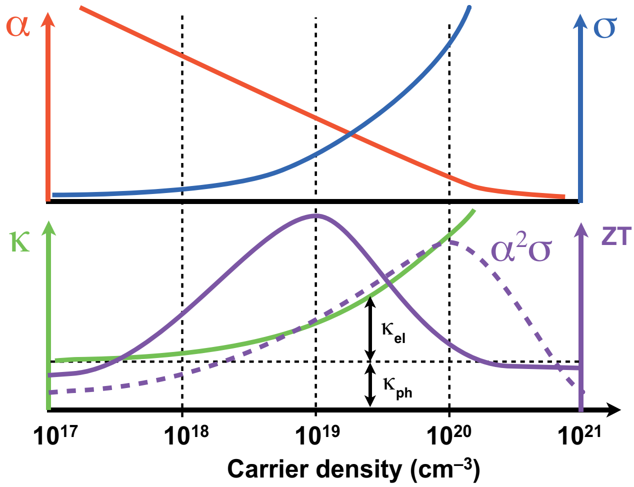


Figure 9. The ZT, Seebeck coefficient, electrical conductivity, thermal conductivity and power factor versus carrier density for a typical semiconductor.

$$\kappa_{el} = \frac{\sigma}{q^2 T}\left[\frac{\langle\tau\rangle\langle E^2\tau\rangle - \langle E\tau\rangle^2}{\langle\tau^3\rangle}\right] \tag{38}$$

The clear result that is not ideal for optimising thermoelectric materials at high carrier densities is that $\kappa_{el} \propto \sigma$.

5. Optimising ZT

In bulk semiconductor thermoelectric materials, once a material with a particular composition has been chosen, the only real parameter that can be varied to optimise ZT is the doping density. The electrical and thermal parameters in bulk materials are coupled through the Wiedemann-Franz law and so simply by improving one parameter through choosing a better doping does not necessarily result in higher ZT. Fig. 9 shows a schematic diagram of the electrical and thermal properties of the bulk thermoelectric semiconductor material Bi_2Te_3 as a function of doping density. For this example, the maximum ZT is close to 10^{19} cm^{-3} whilst the maximum power factor is at 10^{20} cm^{-3}. The figure also demonstrates the coupling of the electrical and thermal conductivities at high carrier densities and the inverse relationship between carrier density and the Seebeck coefficient ($\alpha \propto n^{-\frac{2}{3}}$ where n is the carrier density).

A comparison of the best n-type and p-type ZT values as a function of temperature is demonstrated in Fig. 10. The solid lines are the ZT values for bulk materials. The majority of the 3D bulk results all have ZT values that are around 1 or less. No bulk material has yet been found with a ZT significantly higher than 1. There have, however, been a number of suggestions to improve ZT by going to lower dimensional structures where the

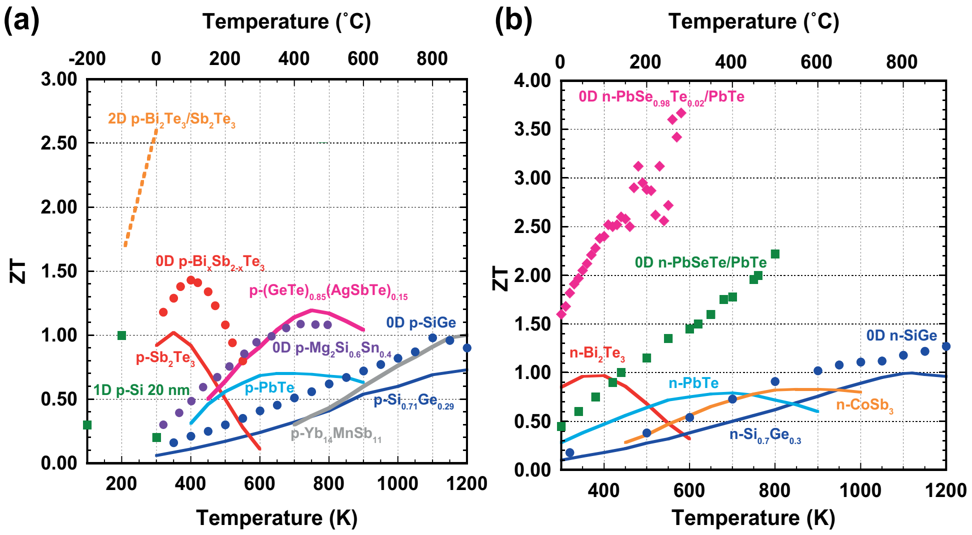


Figure 10. Left: a comparison of ZT for p-type material as a function of temperature (p-Sb$_2$Te$_3$, p-PbTe, p- CeFe$_4$Sb$_{12}$, p-Yb$_{14}$MnSb$_{11}$ [5], p-Si$_{0.71}$Ge$_{0.29}$ [6], 2D p-Bi$_2$Te$_3$/Sb$_2$Te$_3$ [7], 1D Si [8], 0D p-SiGe [9], p-(GeTe)$_{0.85}$(AgSbTe)$_{0.15}$ [10], 0D p-Bi$_x$Sb$_{2-x}$Te$_3$ [11], 0D Mg$_2$Si$_{0.4}$Sn$_{0.6}$ [12]). Right: a comparison of ZT for n-type material as a function of temperature (n-Bi$_2$Te$_3$, n-PbTe, n-CoSb$_3$ [5], n-Si$_{0.7}$Ge$_{0.3}$ [6], 0D PbSeTe [13], 0D n-SiGe [14], 0D n-PbSe$_{0.98}$Te$_{0.02}$/PbTe [15]).

Wiedemann-Franz rule can break down and quantum effects can be used to optimise the ZT value which will be discussed in the next section.

The first major calculations to demonstrate the advantages of moving to low dimensional structures was that of Mildred Dresselhaus [16]. There are multiple ways that low dimensional structures can enhance the value of ZT.

5.1. Low dimensional structures

Before demonstrating the potential enhancements that low dimensional structures can bring to the ZT of a thermoelectric material, the definition of lower dimensional samples must be considered. If a sample is made with dimensions of length, L, width, w and thickness, t then the dimensionality of the system and the appropriate transport regime for electrons or phonons is inferred by comparing the sample dimension to the various scattering lengths and characteristics lengths defined below. Care is always required for a sample can be, for example, the 2D in terms of electrical transport but 3D in terms of thermal transport.

In the Drude model, the electrical conductivity is defined in terms of the elastic scattering time, τ, the effective mass of the electrons in the material, m^* and the carrier density, n as

$$\sigma = \frac{nq^2\tau}{m^*} \tag{39}$$

This equation is very simplistic in terms of the mechanisms which determine the electrical conductivity and dependent on the temperature and material, additional transport and scattering mechanisms including disorder, electron-electron interactions, quantum interference or ballistic transport have to be included. Generally the length scale, l_x associated with a scattering time, τ_x for some scattering process is linked through the diffusion constant, D as

$$l_x = \sqrt{D\tau_x} \tag{40}$$

In this form the mobility is defined as $\mu = \frac{q}{m^*}\tau$ and the Einstein relation relates the mobility to the diffusion constant for an absolute temperature, T as

$$\mu = \frac{qD}{k_B T} \tag{41}$$

In most electronic conduction it is only the electrons close to the Fermi level in energy that need to be considered for which the relevant scale is the Fermi wavelength

$$\lambda_F = \frac{2\pi}{k_F} = \sqrt{\frac{2\pi}{n}} = \frac{h}{\sqrt{2m^* E_F}} \tag{42}$$

The elastic scattering length of electrons is defined as the mean free path, ℓ. (Note - the mean free path is strictly defined as the shortest scattering length between all the scattering mechanisms which includes phase coherent scattering, inelastic scattering, electron-electron scattering, but normally the elastic scattering length is the dominant.) The mean free path is defined generally and for lower dimensions as

$$\begin{aligned} \ell &= v_F\tau = \frac{\hbar k_F}{m^*}\tau && (43)\\ \ell_{3D} &= \frac{\hbar}{m^*}(3\pi^2\frac{n}{g_v})^{\frac{1}{3}}\frac{\mu m^*}{q} = \frac{\hbar\mu}{q}(3\pi^2\frac{n}{g_v})^{\frac{1}{3}} && (44)\\ \ell_{2D} &= \frac{\hbar\mu}{q}\sqrt{2\pi\frac{n}{g_v}} && (45)\\ \ell_{1D} &= \frac{\hbar\mu}{q}\pi\frac{n}{g_v} && (46) \end{aligned}$$

where g_v is the valley degeneracy of the semiconductor. The above equations have assumed that the electrons have a spin degeneracy of 2 which is only untrue under large magnetic fields sufficiently high to split the spin degeneracy.

Moving to thermal transport there are a number of models which can be considered to determine the appropriate length scales. The first is the thermal length, L_T defined as the length over which thermal smearing and the associated phase randomization of an electron of the Fermi distribution which produces and energy uncertainty of $k_B T$. This is given by

$$L_T = \sqrt{\frac{D\hbar}{k_B T}} \tag{47}$$

The phonon group velocity is defined as $\frac{\partial \omega}{\partial q}$ where ω is the phonon angular frequency and q is the phonon wavenumber. The phonon mean free path in the 3D bulk as determined by the Debye theory, which assumes that the phase and group velocity of the phonons are equal, is give by

$$\Lambda_{ph} = \frac{3\kappa_{ph}}{C_v \langle v \rangle \rho} \tag{48}$$

where C_v is the specific heat capacity at constant volume, $\langle v \rangle$ is the average phonon velocity and ρ is the density of phonons. Table **??** gives examples of the phonon mean free path in Si and germanium. The Debye theory is by no means unique and the group velocity of phonons is defined in terms of the phonon dispersion relation as demonstrated by Chen [17]. Table **??** provides a comparison between the mean free paths determined by the Debye and dispersion models. There is a significant difference between the two approaches and this is one of the main issues and problems in determining exactly which dimensionality a system is in terms of the thermal transport. In all the 3D cases below, the numbers are larger than the equivalent electron mean free path.

Material	Model	Specific heat ($\times 10^6$ Jm^{-3}K^{-1})	Group velocity (m s^{-1})	Phonon mean free Λ_{ph} (nm)	Debye temperature (K)
Si	Debye	1.66	6400	40.9	645
Si	Dispersion	0.93	1804	260.4	645
Ge	Debye	1.67	3900	27.5	360
Ge	Dispersion	0.87	1042	198.6	360

Table 1

5.2. Quantum wells

Before describing how low dimensional systems can improve ZT, we first require to find the energy solutions for a low dimensional systems. The simplest solution to investigate is the 2D quantum well for electrons, (Fig. 11) where the approximation is made that the potential energy for electrons inside the quantum, V = 0 and the potential energy outside the well is infinite ($V = \infty$). We require to solve the time independent Schrödinger equation which is

$$-\frac{\hbar^2}{2m^*}\frac{\mathrm{d}^2\psi(z)}{\mathrm{d}z^2} + V(z)\psi(x) = E\psi(z) \tag{49}$$

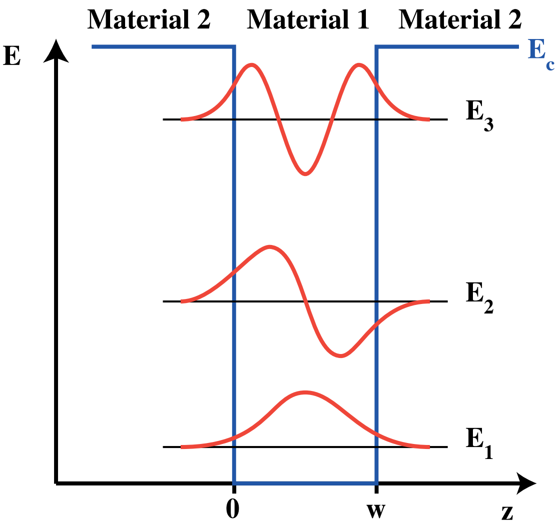


Figure 11. The quantized energy states in a quantum well of width, w.

Outside the quantum well with $V = \infty$, there are no solutions and the electrons are forbidden to occupy regions outside the quantum well. Inside the quantum well $V(z) = 0$ and so the Schrödinger equation reduces to the Helmholtz or wave equation

$$-\frac{\hbar^2}{2m^*}\frac{\mathrm{d}^2\psi(z)}{\mathrm{d}z^2} = E\psi(z) \tag{50}$$

Any travelling wave solution is a valid solution inside the quantum well to this equation so $\psi(z) = A\sin(kz)$, $\psi(x) = A\cos(kz)$, $\psi(z) = Ae^{ikz}$ and $\psi(x) = Ae^{-ikz}$ (where A is an amplitude of the wavefunction) and any mixture of these equations are all potential valid trial solutions. The boundary conditions that a solution must adhere to is that at $z \to \infty$ then $\psi(\pm\infty) = 0$. For the infinite quantum well in Fig. 11 this requires $\psi(0) = \psi(w) = 0$ and so the wavefunctions cannot penetrate outside the quantum well. Also $\psi(z)$ and $\frac{\mathrm{d}\psi}{\mathrm{d}z}$ must be continuous between regions. These boundary conditions therefore require that the only possible solution is

$$\psi(z) = A_n\sin(k_n z) \qquad \text{with } k_n = \frac{n\pi}{w} \text{ and } n = 1, 2, 3, \ldots \tag{51}$$

If this wavefunction is substituted back into the Schrödinger equation then the solution for the energy is

$$E = \frac{\hbar^2 k_n^2}{2m^*} = \frac{\hbar^2\pi^2 n^2}{2m^* w^2} \text{ for } n = 1, 2, 3, \ldots \tag{52}$$

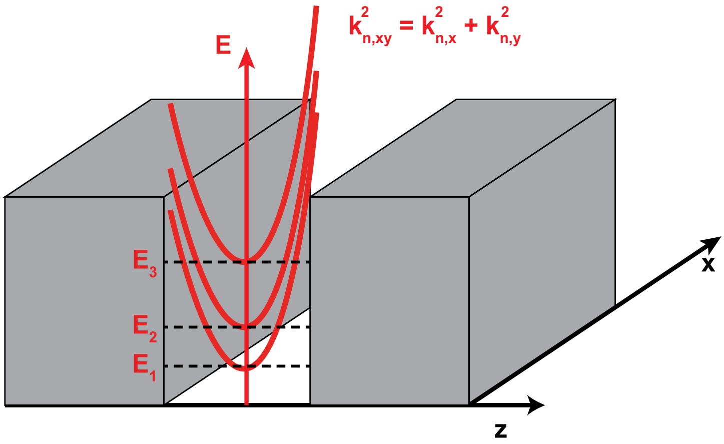


Figure 12. The quantized energy states in a quantum well of width, w.

The integers $n = 1, 2, 3, \ldots$ are called quantum numbers and the solutions provide quantized energy levels that restrict the energy of the electrons into subband states indicated by E_1, E_2, $\ldots E_n$. Whilst in any classical system that has a continuous range of energies the ground state can always have zero energy, in this quantized system however, the ground state with $n = 1$ always has energy. This is known as the zero point energy. In semiconductor materials, if two different materials are placed together then a heterostructure results. By placing one material with a lower conduction band between the same material with a higher conduction band, a quantum well is produced as shown in Fig. 11. The electrons move along the x- and y-directions of the quantum well and are quantized in the z-direction then the electrons have parabolic energy dispersions along the x- and y-directions and the quantized energy dispersion only along the z-direction as shown in Fig. 12. This results in the complete energy of each electron subband to be

$$E = \frac{\hbar^2 k_x^2}{2m^*} + \frac{\hbar^2 k_y^2}{2m^*} + \frac{\hbar^2 \pi^2 n^2}{2m^* w^2} \quad \text{for} \quad n = 1, 2, 3, \ldots \tag{53}$$

5.3. Density of states

Equations 6 and 8 both use the density of states, $g(E)$ which describes the number of available states that electrons can occupy in any material system. The density of states is defined as the number of states per energy per unit volume of real space where if N is the number of states then

$$g(E) = \frac{\mathrm{d}N}{\mathrm{d}E} \tag{54}$$

The density of states is therefore counting the number of states between E and $E + \mathrm{d}E$ in energy. In $\mathbf{k}$-space, the total number of states, N is equal to the volume of the sphere or radius $\mathbf{k}$ first divided by volume occupied by one state and then divided by the volume of real space. Therefore if we have a 3D volume defined by a cube with side length, L then the volume of one state in $\mathbf{k}$-space is $\left(\frac{2\pi}{L}\right)^3$. The number of states is given by

$$N = g_v g_s \frac{4\pi k^3}{3} \frac{1}{\left(\frac{2\pi}{L}\right)^3} \frac{1}{L^3} \tag{55}$$

$$= g_v g_s \frac{4\pi k^3}{3\,(2\pi)^3} \tag{56}$$

where the degeneracy of valleys, g_v and the spin degeneracy g_s have been added. For direct bandgap semiconductors, $g_v = 1$ whilst for indirect bandgap semiconductors the valley degeneracy will be greater than 1. As an example, $g_v = 2$ for silicon. The spin degeneracy, g_s is virtually always 2 for the majority of systems and this only changes for systems with strong magnetic fields. The trick to working out the density of states is to split the derivative in equation 54 into

$$g(E) = \frac{\mathrm{d}N}{\mathrm{d}E} = \frac{\mathrm{d}N}{\mathrm{d}k}\frac{\mathrm{d}k}{\mathrm{d}E} \tag{57}$$

so that equation 56 becomes

$$\frac{\mathrm{d}N}{\mathrm{d}k} = g_v g_s \frac{4\pi k^2}{(2\pi)^3} \tag{58}$$

The parabolic bands of the effective mass theory provide $E = \frac{\hbar^2 k^2}{2m^*}$ which rearranging gives

$$k = \sqrt{\frac{2m^* E}{\hbar^2}} \tag{59}$$

Taking the derivative with respect to energy produces

$$\frac{\mathrm{d}k}{\mathrm{d}E} = \sqrt{\frac{2m^*}{\hbar^2}}\frac{E^{-\frac{1}{2}}}{2} \tag{60}$$

Equations 58 and 60 can now be combined to produce the 3D density of states

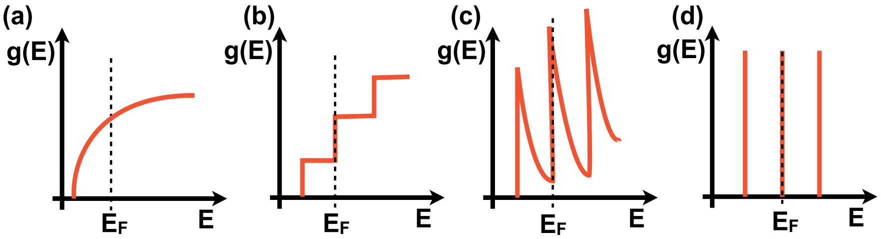


Figure 13. The electron density of states as a function of energy for (a) 3D (b) 2D (c) 1D and (d) 0D semiconductor systems.

$$g_{3D}(E) = \frac{g_v g_s}{4\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{61}$$

The same technique can be repeated for 2D systems where instead of a sphere, the 2D equivalent is a circle in **k**-space. Repeating the technique for the 3D system, the number of states is

$$N_{2D} = g_v g_s \frac{\pi k^2}{3} \frac{1}{\left(\frac{2\pi}{L}\right)^2} \frac{1}{L^2} \tag{62}$$

$$= g_v g_s \frac{\pi k^2}{(2\pi)^2} \tag{63}$$

and repeating the techniques above results in the 2D density of states as

$$g_{2D}(E) = g_v g_s \frac{m^*}{2\pi\hbar^2} \tag{64}$$

This results is for a single subband in a quantum well (see Figs. 11 and 12) and for a heavily doped system as most thermoelectric materials are requires a summation of the density of states over all the subbands. This results in the 2D density of states at an energy, E being the sum over all subband below that energy which is

$$g_{2D}(E) = \sum_{i=1}^{n} g_v g_s \frac{m^*}{2\pi\hbar^2}\Theta\left(E - E_i\right) \tag{65}$$

where Θ is the Heaviside step function.

The technique can be repeated for 1D and 0D systems and to summarise the density of states as a function of dimension are

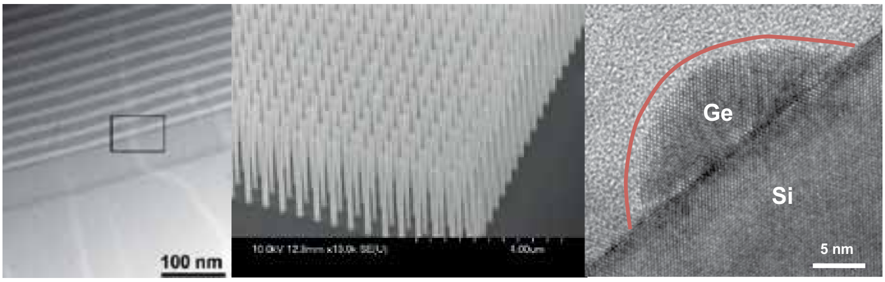


Figure 14. Left: A cross sectional TEM image of Ge quantum wells with $Si_{0.2}Ge_{0.8}$ barriers forming a 2D thermoelectric system [18]. The square highlights a dislocation that limits the performance of this material. Middle: A SEM image of etched 50 nm wide nanowires of Ge/$Si_{0.2}Ge_{0.8}$ material forming 1D thermoelectric systems. Right: A TEM image of a Ge quantum dot grown on a silicon substrate forming a 0D thermoelectric system for scattering phonons.

$$g_{3D}(E) = \frac{g_v g_s}{4\pi^2}\left(\frac{2m^*}{\hbar^2}\right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{66}$$

$$g_{2D}(E) = \sum_{i=1}^{n} g_v g_s \frac{m^*}{2\pi\hbar^2}\Theta\left(E - E_i\right) \tag{67}$$

$$g_{1D}(E) = \frac{1}{2\pi\hbar}\sum_{i=1}^{n} g_v g_s \sqrt{\frac{m^*}{2}}\pi\hbar^2\Theta\left(E - E_i\right) \tag{68}$$

$$g_{0D}(E) = \sum_{i=1}^{n} g_v g_s \delta\left(E - E_i\right) \tag{69}$$

with $\delta(E - E_i)$ the Dirac delta function.

5.4. Low dimensional enhancements to ZT

The density of states as a function of energy for 3D, 2D, 1D and 0D are plotted in Fig. 13. Also plotted is the ideal position for the Fermi energy if the Seebeck coefficient is to be maximised using the Cutler and Mott equation

$$\alpha = -\frac{\pi^2}{3q}k_B^2 T\left[\frac{d\ln(\mu(E)g(E))}{dE}\right]\Bigg|_{E=E_F} \tag{70}$$

By moving to lower dimensional structures, there is a larger asymmetry in the density of states around the Fermi energy and the above equation and the discussions in section 2 indicates that this increases the Seebeck coefficient. Therefore by choosing systems with lower dimensions, the Seebeck coefficient can be enhanced. A number of experimental examples have demonstrated significant improvements to the Seebeck coefficient from reducing the dimensionality from 3D to 2D, 1D or 0D. Examples of the physical structures are shown in Fig. 14. Quantum wells with the transport either parallel or perpendicular to the quantum wells are used for 2D thermoelectric systems (see Fig. 14 left). Nanowires either grown or etched can be used to form 1D thermoelectric systems (Fig. 14 middle) whilst most 0D thermoelectric systems use quantum dots that are aimed at scattering phonons (Fig. 14 right).

2D quantum well systems can also enhance the electrical conductivity by a number of techniques. If the electrical and thermal transport is along quantum wells then modulation doping can be used to enhance the electrical conductivity. Here the dopants are only placed in the barriers which are at higher energy than the quantum well. The carriers fall into the quantum well and are therefore remote from the ionised dopants that created the carriers. By separating the carriers from the dopants, the Coulomb scattering is reduced and the mobility and electrical conductivity increases. This occurs along with the Seebeck enhancement described above and so higher power factors can be created. The potential disadvantage of this technique is that the electrical conductivity can be so large that it also can increase the thermal conductivity in a detrimental way to the ZT. By optimising the parameters, higher ZTs can be produced.

The third approach to optimising ZT is to reduce the thermal conductivity. There are many ways to reduce the thermal conductivity by adding scattering centres or rough interfaces that scatter phonons. The main issue is to aim for a scattering technique that scatters phonons more than electrons. This is not as easy as it sounds. Adding 0D nanoparticles or quantum dots into a material has been successful at reducing κ faster than σ as shown in Fig. 10 in a number of material systems for both n- and p-type semiconductors [13, 15]. The quantum dots when of the correct size can scatter phonons much more easily than electrons especially for highly doped samples where the electron mean free paths are typically longer than phonon mean free paths. There are a number of examples of materials such as skutterudites where heavy atoms are inserted to fill voids in the lattice to have a similar effect at the microscale of the lattice [5].

2D superlattices with the electron transport perpendicular to the quantum well and barriers are also good at scattering phonons. The disadvantage of this type of superlattice is that the electrons or holes must quantum mechanically tunnel through the barriers which also significantly reduces the electrical conductivity to typically 3 to 4 times lower than bulk material. The lower thermal conductivity combined with the higher Seebeck from the 2D quantum wells does produce significant enhancements to ZT as show in Fig. 10 [7].

Finally 1D nanowires have also demonstrated substantial improvements to ZT. Boukai et al., [8] demonstrated 10 nm wide Si nanowires which demonstrate enhanced Seebeck coefficients and significantly reduced thermal conductivities compared to bulk Si. The thermal conductivity demonstrated the largest changes with reductions of up to 150 times that of bulk silicon whilst the Seebeck improves by a factor of 2. ZTs at room temperature of 0.3 have been achieved with higher values at lower temperatures. The nanowires demonstrate how confining phonons in low dimensional structures can make significant changes to the ZT of a material.

6. The power output from thermoelectric modules

To this point the majority of this review has concentrated on ways to optimise ZT but for applications it is the current and voltage i.e. the output power produced that is the most

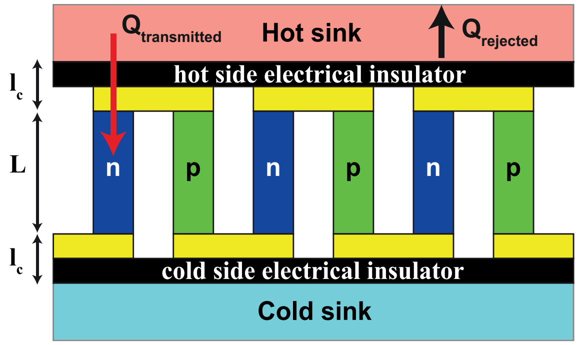


Figure 15. A schematic diagram of a complete module with leg length, L and contact length, l_c.

important parameter for applications. Therefore it is important to understand the issues of the output power and how it may be optimised. We will start by considering a complete module as shown in Fig. 15. As the Seebeck coefficient of most materials is $\ll 1$ mV/K and most applications will require at least 1 V and in many cases multiple Volts, a large number of legs must be connected in series to achieve suitable output voltage for the applications. If we have N legs each of length, L, with thermal conductivity, κ and contacts of length, l_c with thermal conductivity, κ_c then the voltage produced is given by [19]

$$V = \frac{\alpha N \Delta T}{1 + 2\frac{\kappa l_c}{\kappa_c L}} \tag{71}$$

The current can also be calculated and for legs of area A and electrical conductivity, σ with specific contact resistivity, ρ_c, the current is

$$I = \frac{\alpha \sigma A \Delta T}{2(\rho_c \sigma + L)(1 + 2\frac{\kappa l_c}{\kappa_c L})} \tag{72}$$

By multiplying the voltage and current together, the resulting power is

$$P = \frac{\alpha^2 \sigma A N \Delta T^2}{2(\rho_c \sigma + L)(1 + 2\frac{\kappa l_c}{\kappa_c L})^2} \tag{73}$$

There are a number of issues that the power equation highlights. The first is that the power is dominated by $\alpha^2\sigma$ which is called the power factor. The second is that the power is proportional to the area of the legs and the number of the legs in the module. The power is also proportional to the square of ΔT. Finally, whilst shortening the length of the legs in the module to first order will increase the power, equation 73 demonstrates that as the leg length reduces, the contact resistance of each leg plays a larger part in reducing the output power. For the microfabricated modules, having a low specific contact resistivity can be as important as a material with high ZT and power factor in being able to achieve a high output power.

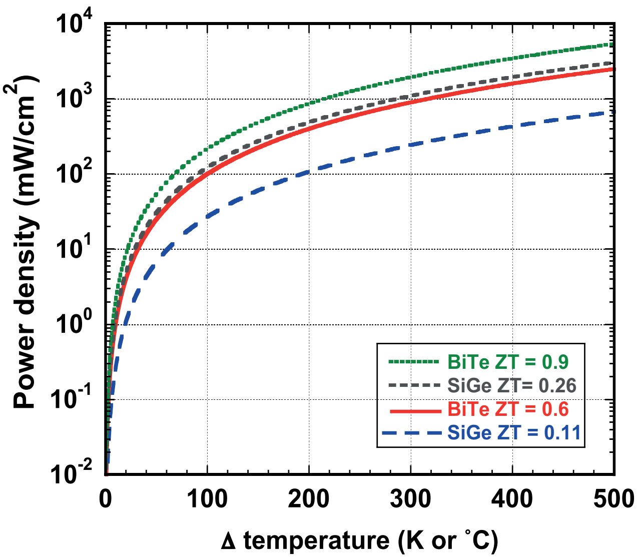


Figure 16. The power as calculated from equation 73 for a module with 525 legs of area 500 μm × 50 μm, leg length of L = 20 μm and l_c = 10 nm.

Another issue is that for many applications the amount of heat may vary and is not constant. Therefore power conditioning along with electrical impedance matching is required so that the thermoelectric module can always operate at the maximum power point. Maximum power point tracking systems are required which automatically adjust the load impedance so that it is always matched to the thermoelectric for all levels of heat and ΔT being applied across the thermoelectric.

Material	$\alpha_p - \alpha_n$	σ	κ	ZT	ρ_c
Units	(μV/K)	(S/m)	(W/mK)	@ 300 K	Ω-cm^2
BiTe	500	71468	2	0.67	10^{-7}
BiTe	500	71468	2	0.67	10^{-7}
SiGe	600	70000	7.3	0.26	10^{-8}
SiGe	500	27900	4.8	0.22	10^{-8}

Table 2

7. Applications

The first thermoelectric application was to power satellites for space applications in 1961 [20]. Space systems use radioisotope thermoelectric generators (RTGs) where a radioactive material heated by the decay and emission of radiation is used as the hot source with a thermoelectric generator to turn the heat into electricity. Plutonium 238 was the main power source used by NASA in most of their 28 RTG systems which operates with temperatures up to 1000 ^{o}C whilst the outside of the spacecraft is used with heat exchangers to provide the cold sink. With such high temperatures, SiGe has been the main thermoelectric material

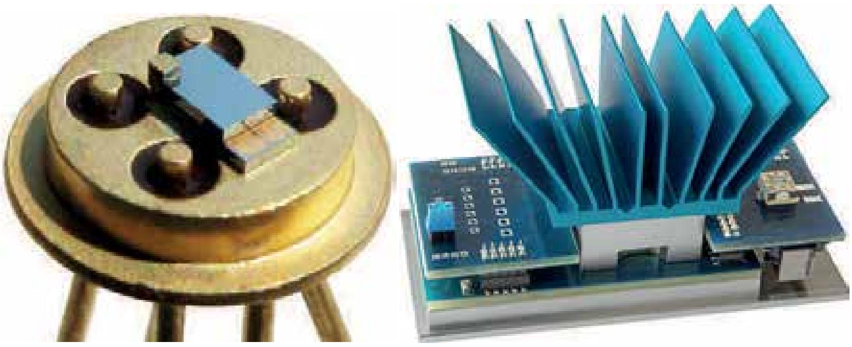


Figure 17. Left: a telecoms laser on a microfabricated Peltier cooler produced by Micropelt. Right: a thermoelectric generator produced by Micropelt showing the cold sink aimed at dissipating heat through air cooling. Copyright Micropelt [21].

used for these generators and the efficiencies can be as high as 6.6 % mainly due to the high ΔT. Both Voyager space missions which are now outside the solar system are powered by SiGe RTGs weighing 37.7 kg and provided 470 W on launch at 40 V. Over a period of time the temperature of the plutonium reduces as a function of the half-life of the radioactive decay so the systems are now generating less than 350 W, 34 years after launch. This is one of the best demonstrations of the robustness of the thermoelectric generator as for such space systems now over a light year away from earth, the power sources must be fit and forget. One problem is that there are no present sources of plutonium 238 as all the weapons nuclear reactors have been shut down and so there are a number of research programmes aiming to develop new RTG technology using available radio isotopes.

The major application for thermoelectric devices at present is as Peltier coolers (Fig. 17 left) to maintain a low temperature for electronic and optoelectronic components such as telecoms lasers and rf sources. It is therefore in the use as a cooler rather than generating electricity that thermoelectrics are predominantly used at present. A number of companies are offering thermoelectric generator demonstrator kits as products (Fig. 17 right) to allow companies to test thermoelectrics for specific applications. The major present applications are for generating electricity to run sensors in a range of predominantly industrial environments where ΔTs between 20 and 100 ^{o}C can provide sufficient energy for the sensors. The majority of the energy use in the sensors is for the wireless communication that for mobile phone system communication will require powers around 5 mW while many of the sensing elements only require sub-μW powers for making a measurement. These wireless industrial sensing applications are widespread and the cost of the thermoelectric device becomes economical for systems where no installation of wires for communication and/or long term power. It is the large cost of replacing batteries (mainly labour costs) that allows thermoelectrics and other forms of energy harvesting to be cost effect. Most wireless sensors systems now require between 1 and 5 mW of power to run mainly dependent on the distance for the communication and so a square cm area thermoelectric device requires around 50 ^{o}C to provide sufficient power. As the communications consumes the most power, most systems have rechargable battery or super-capacitor storage systems and then use burst modes of communication so that information is only sent when required to save power.

Research results have been published on clothing with integrated thermoelectrics [22]. Only small ΔTs can be provided across clothing using the air temperature outside the

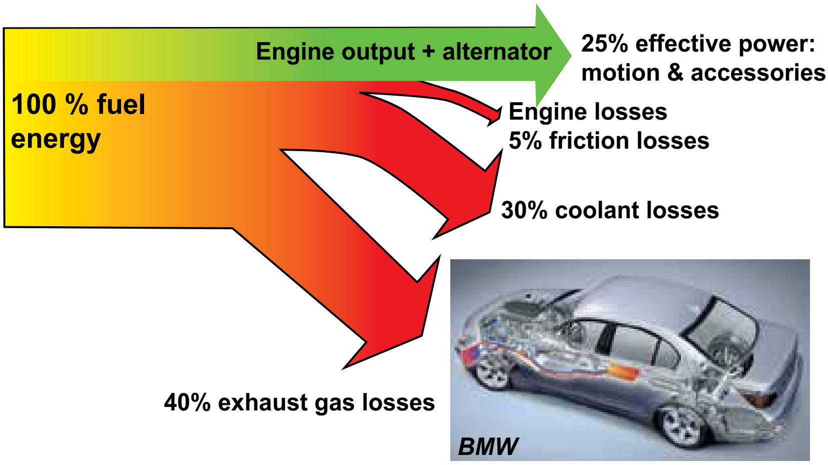


Figure 18. A schematic diagram of how the energy from a combustion engine in a car is distributed. 25% of the energy produces motion and through the alternator generates electricity to power accessories including the electrics, the air conditioning and the hifi system. 75% of the energy from the fuel is lost mostly through friction and heat. 40% of the fuel energy disappears through the exhaust system hence there is interest in using thermoelectrics to harvest some of this waste energy.

body to provide the cold side temperature. There is enormous interest in such power sources for autonomous wireless health monitoring systems that can be fit and forget. Electrocardiography systems integrated into the clothing and powered by thermoelectrics have already been demonstrated and even tested through cycles of washing in washing machines to check the robustness of the technology. For this technology to be practical, batteries or super capacitors and power management must also be integrated along with the sensors, some processing and communication technology. Whilst the low powers available with the small ΔTs may struggle to power the present communication systems, new short range communication protocols being developed at present for autonomous sensors by the IEEE are required before this type of application can be aggressively pursued.

The major driver for improved thermoelectrics at present is probably the car industry where European legislation to improve fuel efficiency is driving thermoelectrics research to replace the alternator. The car is an excellent system where thermoelectrics could play a big role as 75% of the fuel ends up as waste heat and the 40 % of waste heat that goes down the exhaust pipe into an environment that could be used to capture this heat and convert it into electricity (Fig. 18) [23, 24]. The temperature of the exhaust system can range from room temperature up to 750 ^{o}C so this is driving work on new thermoelectric materials to replace the best at present which is PbTe with toxic Pb that cannot be used for applications. Initial modelling has suggested that up to a 5% in fuel consumption could be achieved with suitable thermoelectrics with ZT of 1 but the key issue is getting the whole thermoelectric system cheap enough for the market. Also no thermoelectric provides ZT of 1 from room temperature to 750 ^{o}C so segmented modules and/or new materials are required. Most of the major car companies are now working heavily of thermoelectrics and it is only a mater of time before automotive systems become available.

A developing application is thermal photovoltaics. Concentrator photovoltaics are now producing efficiencies of around 45 % when the light is concentrated up to a thousand times. These systems are aimed at solar farms which in sunny regions of the earth have the potential to produce power stations at the 100s of MW scale. Such concentration, however, results in the photovoltaic cells being heating to very high temperatures that results in large thermal cycling which ultimately produces failures reducing the lifetime of the systems. By integrating a thermoelectric with the photovoltaic, not only can electricity be generated from the heat thereby providing a joint and therefore improved system efficiency closer to the Carnot limit but more importantly, the thermoelectric cools the photovoltaic reducing the thermal cycling extremes and increases the lifetime of the system. This increase in lifetime results in substantially cheaper cost per Watt which is the major driver for photovoltaics. It is clear this will be a developing application for thermoelectrics in the future.

At present thermoelectrics requires a "killer application" before volume manufacture will result in widespread use. The automotive applications at present appear to be the major application driver, more by legislation rather than market but it is clear that the potential requirements for thermoelectric generators will improve as fossil fuel prices increase in the future. The autonomous sensor market may well also drive thermoelectrics but the real problem that must be solved is to find sustainable thermoelectric materials. Tellurium is the 9^{th} rarest element on earth predicted to run out before 2020 with the present use and so a key requirement for future thermoelectrics is that Te-free materials can be found that can be cheaply produced and with a high ZT and power factor. There is presently an enormous amount of research in thermoelectrics especially aiming to find Te-free materials and to produce modules for higher temperature applications such as the automotive energy harvesting. For students starting in a research career, this is one research area that is likely to expand over the next decades.

Author details

Douglas Paul

University of Glasgow, School of Engineering, Rankine Building, U.K.

References

[1] J.M. Ziman. *Electrons and Phonons*. Qxford University Press, 1960.

[2] Melvin Cutler and N. F. Mott. Observation of anderson localization in an electron gas. *Phys. Rev.*, 181(3):1336–1340, May 1969.

[3] J. Callaway. Model for lattice thermal conductivity at low temperatures. *Phys. Rev.*, 113(4):1046 – 1051, 1959.

[4] B.R. Nag. *Electron Transport in Compound Semiconductors*, volume 11 of *Solid State Sciences*. Springer, 1980.

[5] G. J. Snyder and E. S. Toberer. Complex thermoelectric materials. *Nature Materials*, 7(2):105–114, 2008.

[6] J.P. Dismukes, E. Ekstrom, D.S. Beers, E.F. Steigmeier, and I. Kudman. Thermal + electrical properties of heavily doped ge-si alloys up to 1300 degrees k. *J. Appl. Phys.*, 35(10):2899, 1964.

[7] R. Venkatasubramanian, E. Siivola, T. Colpitts, and B. O'Quinn. Thin-film thermoelectric devices with high room-temperature figures of merit. *Nature*, 413(6856):597–602, 2001.

[8] A. I. Boukai, Y. Bunimovich, J. Tahir-Kheli, J. K. Yu, W. A. Goddard, and J. R. Heath. Silicon nanowires as efficient thermoelectric materials. *Nature*, 451(7175):168–171, 2008.

[9] G. Joshi, H. Lee, Y. C. Lan, X. W. Wang, G. H. Zhu, D. Z. Wang, R. W. Gould, D. C. Cuff, M. Y. Tang, M. S. Dresselhaus, G. Chen, and Z. F. Ren. Enhanced thermoelectric figure-of-merit in nanostructured p-type silicon germanium bulk alloys. *Nano Letters*, 8(12):4670–4674, Dec 2008.

[10] D.M. Rowe, editor. *Thermoelectrics Handbook: Micro to Nano*, Boca Raton, FL, USA, 2006. CRC Press, Taylor and Francis.

[11] Yi Ma, Qing Hao, Bed Poudel, Yucheng Lan, Bo Yu, Dezhi Wang, Gang Chen, and Zhifeng Ren. Enhanced thermoelectric figure-of-merit in p-type nanostructured bismuth antimony tellurium alloys made from elemental chunks. *Nano Letters*, 8(8):2580–2584, 2008. PMID: 18624384.

[12] Q. Zhang, J. He, T. J. Zhu, S. N. Zhang, X. B. Zhao, and T. M. Tritt. High figures of merit and natural nanostructures in mg[sub 2]si[sub 0.4]sn[sub 0.6] based thermoelectric materials. *Applied Physics Letters*, 93(10):102109, 2008.

[13] T. C. Harman, P. J. Taylor, M. P. Walsh, and B. E. LaForge. Quantum dot superlattice thermoelectric materials and devices. *Science*, 297(5590):2229–2232, Sep 2002.

[14] X. W. Wang, H. Lee, Y. C. Lan, G. H. Zhu, G. Joshi, D. Z. Wang, J. Yang, A. J. Muto, M. Y. Tang, J. Klatsky, S. Song, M. S. Dresselhaus, G. Chen, and Z. F. Ren. Enhanced thermoelectric figure of merit in nanostructured n-type silicon germanium bulk alloy. *Appl. Phys. Lett.*, 93(19):193121, Nov 2008.

[15] T.C. Harman, M.P. Walsh, B.E. laforge, and G.W. Turner. Nanostructured thermoelectric materials. *J. Electronic Materials*, 34:L19–L22, 2005.

[16] L. D. Hicks and M. S. Dresselhaus. Effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B*, 47(19):12727–12731, 1993.

[17] G. Chen. Thermal conductivity and ballistic-phonon transport in the cross-plane direction of superlattices. *Phys. Rev. B*, 57(23):14958–14973, Jun 1998.

[18] A. Samarelli, L.Ferre Llin, S. Cecchi, J. Frigerio, T. Etzelstorfer, E.Müller Gubler, Y. Zhang, J. R. Watling, D. Chrastina, G. Isella, J. P. Hague, J. Stangl, J.M.R. Weaver, P. S. Dobson, and D.J. Paul. The thermoelectric properties of ge/sige modulation doped superlattices. *J. Appl. Phys.*, 113:233704, 2013.

[19] D.W. Rowe and G. Min. Design theory of thermoelectric modules for electrical power generation. *Science, Measurement and Technology, IEE Proceedings -*, 143(6):351–356, 1996.

[20] R.D Abelson. Space missions and applications. In D.M. Rowe, editor, *Thermoelectric Handbook: Macro to Nano*, chapter 56, pages 56–1 – 56–29. Taylor and Francis, 2006.

[21] 2013.

[22] Vladimir Leonov, Tom Torfs, Chris Van Hoof, and Ruud J. M. Vullers. Smart wireless sensors integrated in clothing: an electrocardiography system in a shirt powered using human body heat. *Sensors & Transducers*, 107(8):165 – 176, 2009.

[23] K Matsubara and M Matsuura. A thermoelectric application to vehicles. In D.M. Rowe, editor, *Thermoelectric Handbook: Macro to Nano*, chapter 52, pages 52–1 – 52–11. Taylor and Francis, 2006.

[24] Jihui Yang and Francis R. Stabler. Automotive applications of thermoelectric materials. *Journal of Electronic Materials*, 38(7):1245 – 1251, 2009.

Chapter 5

Electromagnetic Radiation Energy Harvesting – The Rectenna Based Approach

Gabriel Abadal, Javier Alda and Jordi Agustí

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57118

1. Introduction

The energy available in the electromagnetic spectrum

How much energy is available around us? Which use can we give to this energy? These are two questions to which answers had been changing over time. What would be our particular answer if a forefather or an ancestor would ask them to us? Some sources of energy like sun, wind or sea waves have been present unaltered since the prehistoric times and before to nowadays. Some others like oil and natural gas have been progressively reduced by the action of man. But it is interesting to notice that there are some other sources, which we can name as artificial sources, and that have emerged by man's action, as a consequence of industrial and technological development. Such modern or artificial energy sources are directly connected to the energy harvesting technology since, for instance, most of the vibrations or temperature gradients are produced by machines and engines. Also in the electromagnetic spectrum, we can harvest energy not only from natural sun radiation, but also from all the artificial radio-frequency sources that are permanently increasing in number and which are a consequence of one of the last technological revolutions: the Information and Communications Technology (ICT) revolution.

Although when we think about electromagnetic (EM) waves at present time, we probably tend to think about examples like radio, TV or cell phones, where the information part of the electromagnetic signal is the protagonist, we should bear in mind that those signals are in fact a combination of information and energy. In this chapter, we are not interested in describing how information can be transmitted through electromagnetic waves but how the energy of these waves is transmitted and collected to be harvested and used to supply ICT devices. In order to calculate how much energy can be associated to an electromagnetic wave, we have to consider the physical nature of these particular waves.

1.1. Basic concepts

Electromagnetic waves in the electromagnetic spectrum (figure 1) are characterized by their wavelength λ or, alternatively, by their frequency ν. Both magnitudes are related with the propagation speed of such waves, the speed of light c, through:

$$c=\lambda \cdot \nu \tag{1}$$

On the other hand, the frequency of EM radiation is directly related to the energy E of a photon associated to this radiation, i.e. the quantum of EM radiation or the most fundamental constitutive part of this radiation as defined by quantum mechanics, by

$$E=h \cdot \nu \tag{2}$$

where h=6.626 10^{-34} J s is the Planck constant.

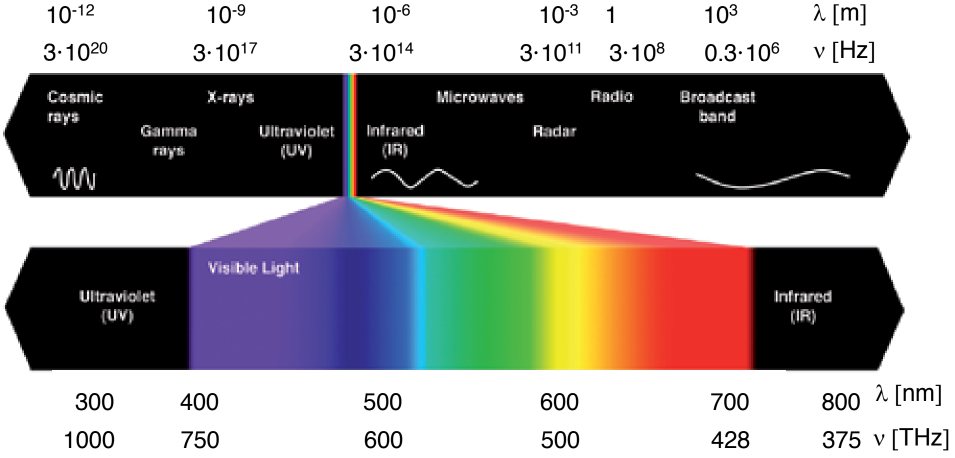
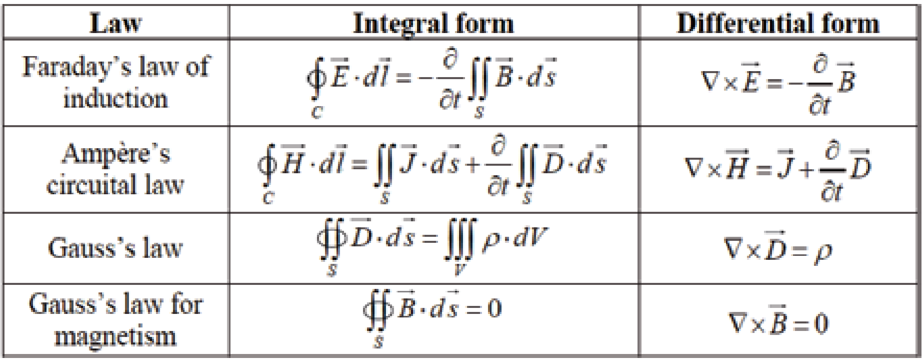


Figure 1. Diagram of the electromagnetic spectrum with indications of the wavelength, λ, and frequency, ν, of the most representative radiations from shorter and most energetic, cosmic rays, to the longer and less energetic radiofrequencies. A zoom detail of the optical part of the spectrum shows that light radiations is in the hundred nm and THz range of wavelengths and frequencies respectively.

Unlike what occurs in photovoltaics technology, where optical radiation energy is better accounted in terms of photon energy since there the conversion mechanism is based in photon-electron interactions, in rectenna technology it is more convenient to express the input EM radiation in terms of the power or the power density of the EM wave.

An EM wave can be defined as a form of energy radiated by a source which results in a combination of oscillating electric and magnetic fields. In most of materials, the direction of the EM wave propagation is perpendicular to the electric and magnetic fields, which are also oscillating in phase perpendicular to each other.

The set of equations which describe how electric and magnetic fields propagate, interact and how they are influenced by material properties are Maxwell's equations. An EM wave can be described with these equations, which must be met for a set of particular boundary conditions. Maxwell's equations are summarized in Table 1.



Law	Integral form	Differential form
Faraday's law of induction	$\oint_{C} \vec{E} \cdot d\vec{l} = -\frac{\partial}{\partial t} \iint_{S} \vec{B} \cdot d\vec{s}$	$\nabla \times \vec{E} = -\frac{\partial}{\partial t} \vec{B}$
Ampère's circuital law	$\oint_{C} \vec{H} \cdot d\vec{l} = \iint_{S} \vec{J} \cdot d\vec{s} + \frac{\partial}{\partial t} \iint_{S} \vec{D} \cdot d\vec{s}$	$\nabla \times \vec{H} = \vec{J} + \frac{\partial}{\partial t} \vec{D}$
Gauss's law	$\oiint_{S} \vec{D} \cdot d\vec{s} = \iiint_{V} \rho \cdot dV$	$\nabla \times \vec{D} = \rho$
Gauss's law for magnetism	$\oiint_{S} \vec{B} \cdot d\vec{s} = 0$	$\nabla \times \vec{B} = 0$

Table 1. Maxwell's equations.

where $\vec{E}$ is the electric field intensity, $\vec{B}$ and $\vec{H}$ are the magnetic fields, $\vec{J}$ is the total current density, $\vec{D}$is the electric displacement field and ρ is the total charge density.

The propagation of a plane EM wave can be described by the EM wave equation, which can be derived from Maxwell's equations. The homogeneous form of this second-order differential equation can be written in terms of either the electric field or the magnetic field as

$$\left(\nabla^2 - \mu\varepsilon \frac{\partial^2}{\partial t^2}\right)\begin{Bmatrix} E_y(x,t) \\ B_z(x,t) \end{Bmatrix} = 0 \tag{3}$$

where μ and ε are the permeability and the permittivity of the propagation medium, respectively.

Knowing that EM waves carry energy with them in the form of electric and magnetic fields, we can compute their energy flow per unit area using the so called Poynting vector

$$\vec{S} = \vec{E} \times \vec{H} \tag{4}$$

From the Poynting vector and considering a uniform plane wave the time-average power density of the EM wave can be computed as [1]

$$P_{av} = \frac{1}{2} \cdot \frac{|E_0|^2}{\operatorname{Re}\{\eta\}} \tag{5}$$

where E_0 is the peak value of the electric field and η is the impedance of the propagating medium. If the wave propagates in a loss-less dielectric medium η is a real number. Being this medium the free space, the impedance can be computed as follows:

$$\eta_0 = \sqrt{\frac{\varepsilon_0}{\mu_0}} = \frac{1}{\varepsilon_0 \cdot c_0} \tag{6}$$

where μ_0 is the vacuum permeability, ε_0 is the vacuum permittivity and c_0 is the speed of light in free space. The value for the impedance of the vacuum is about 377Ω.

A good approximation to the radiated power at a certain distance d from an emitter can be computed considering that the emitter is an isotropic radiator (EM point source which radiates the same power in all directions)

$$\mathrm{P}_{\mathrm{r}d} = \frac{P_{rT}}{4 \cdot \pi \cdot d^2} \tag{7}$$

where P_{rT} is the total radiated power and d the distance from the emitter. Notice that real antennas do not radiate isotropically, they have a certain radiation pattern which depends mainly on the geometry of the antenna and the surrounding media.

In 1999 the Council of the European Union made some recommendations on the limitation of exposure to electromagnetic fields [2]. Table 2 summarizes the maximum recommended values for the electric field.

Frequency range	E-field strength (V/m)
0-1 Hz	–
1-8 Hz	10000
8-25 Hz	10000
0.025-0.8 kHz	250 / f
0.8-3 kHz	250 / f
3-150 kHz	87
0.15-1 MHz	87
1-10 MHz	87 / $f^{0,5}$
10-400 MHz	28
400-2000 MHz	1,375 · $f^{0,5}$
2-300 GHz	61

Table 2. Reference levels for electric fields from 0 Hz to 300 GHz.

On the other hand, the IEEE International Committee on Electromagnetic Safety has made some additional recommendations in order to protect human beings from harmful effects caused by the exposure to electromagnetic fields [3]. Table 3 summarizes the maximum recommended values for the RMS electric field, magnetic field and power density.

Frequency range (MHz)	RMS electric field strength (V/m)	RMS magnetic field strength (A/m)	RMS power density (E-field, H-field) (W/m^2)
0.1-1	1842	16,3 / f	(9000 , 100000 / f^2)
1-30	1842 / f	16,3 / f	(9000 / f^2 , 100000 / f^2)
30-100	61,4	16,3 / f	(10 , 100000 / f^2)
100-300	61,4	0,163	10
300-3000	–	–	f / 30
3000-30000	–	–	100
30000-300000	–	–	100

Table 3. Reference levels for electric, magnetic fields from 0 Hz to 300 GHz.

Finally, although tables 2 and 3 give a good idea of the maximum energy available from RF emissions in terms of electric field and power density, in table 4 power density values and ranges corresponding to different applications are also summarized and compared to sunlight in the visible range.

Application	Power density (mW/cm^2)
Old UHF TV band	10^{-9}
FM radio @ 50 km from 100kW base station	10^{-7}
ISM bands: Zigbee/Bluetooth/WIFI	$10^{-8}/10^{-7}/10^{-6}$
Standard ambient level with no high power equipment	10^{-6} – 10^{-5}
GSM, UMTS (3G telecom) @ 10 m from base station	10^{-6} – 10^{-4}
Cellular phone @ 50 m from base station	10^{-4} – 10^{-2}
Solar Power Satellite (SPS) Wireless Power Transmission (WPT)	10^{-1} - 10
Solar radiation in the visible range	10^{2}

Table 4. Comparison of power densities for different applications with solar radiation in the visible range.

1.2. Photovoltaics versus rectenna technologies

When electromagnetic waves were experimentally observed, they were generated using antennas and radiating elements. Along the development of radio emission, antenna design became a separate area of expertise where the geometry of those elements configured the characteristics and capabilities of emission and reception of the EM waves. The shape and orientation of those antennas determine the polarization and direction of the emission, and reception. Electromagnetic spectrum was mastered and used in science and technology. Fortunately, the wavelengths associated with the radioelectric and microwave spectra allowed the manufacturing of radiating elements with the available fabrication tools. When increasing the frequency of the electromagnetic radiation, the geometries were shrunk accordingly and new fabrication strategies were used. Actually, an important leap in antenna design and fabrication appeared when using planar antennas written on flat substrates by microlithography techniques. Millimeter waves and Terahertz still benefit from those fabrication techniques. However, when the optical domain was placed as a feasible goal for antenna design, the use of electron beam lithography, focused ion beam, and related nanometric precision manufacturing tools were necessary. Even more, those metals traditionally used as materials for antenna fabrication appeared to behave as non-perfect conductors, showing spectral dispersion and a non-negligible penetration depth.

At the same time that antennas were clearly devoted to the emission and detection of EM wave in the radioelectric, and microwave regimes, light and optical spectrum was covered with other reliable technologies for emission (incandescence lamps, spectral lamps, lasers, etc.) and detection (Golay cells, thermoconductors, photovoltaics, etc.) mainly based for detection in the quantified energy levels of semiconductors. Then, photodetectors improved their performance in responsivity, signal-to-noise ratio, cut-off frequency, size, and biasing requirements.

Then, it is easy to understand that antennas did not find a suitable place to develop as optical detectors. Semiconductor detectors were here to stay, fabrication of optical antennas is difficult and requires high-tech machinery, and metals are non-perfect conductors anymore in the optical regime.

However, some advances were made in using antenna-coupled detectors in the detection of light at higher and higher frequencies, and in its use as frequency mixer or coupled to bolometric devices. Besides, nanoscience has found optical antennas as promising elements to explore materials and media with high spatial resolution. Plasmonic optics has become an emerging field, where the collective oscillation of charges produces exotic phenomenologies that are used for sensing and probing sub-wavelength structures.

Several reports and papers [4,5,6,7] have been published in the past years presenting optical antennas and rectennas as harvesters of electromagnetic radiation in the infrared and visible spectrum. They are based on the principle of rectification of the currents generated in an antenna structure that resonates at the visible frequency. The idea, although appealing, has been somehow over-estimated when promising efficiencies above 80%. However, as we will see in this chapter, some important problems need to be addressed before fabricating an operative device. Unfortunately, the task of rectifying electric fields oscillating at 10^{14}-10^{15}Hz frequencies is formidable, and the efficiency figures obtained so far are well below the announced limit. The bottleneck of the technology remains in the rectification process. At the same time, some important advances have been made to tailor the impedance of optical antennas to properly couple the electromagnetic field and also to transfer the power to the load, i.e., to the rectifier. Then, optical rectennas can be considered as a promising technology with high potential. Based on the current results, more effort needs to be allocated to leap over the rectifying mechanism with novel technologies.

Although it is limited to the solar region of the electromagnetic spectrum, the most mature and standard technology (developed since the mid 70's) to harvest energy from EM radiation is photovoltaics (PV). According to the *National Renewable Energy Laboratory (NREL)*, conversion efficiency of PV technologies has been increasingly evolved during the last 40 years (figure 2). From the most simple variant of the 1st generation represented by the silicon based cells, to the 2nd and 3rd generations corresponding to thin-film and the most sophisticated multijunction cells respectively, a trade-off between efficiency and production cost is defining the market of each variant (table 5).

PV technology	Efficiency (%)	Market Share (%)
1st generation	20	90
2nd generation	5-12	10
3rd generation	40-50	--

Table 5. Efficiency versus market of the 3 different PV technology generations.

The basic element in PV technology is the photovoltaic/solar panel/module which is composed of photovoltaic cells connected in parallel, when photogenerated current must be enhanced, or placed in series, when the output voltage is the parameter that needs to be maximized (see chapter 10: "Electronics for Power and Energy Management").

The working principle of a photovoltaic cell is based on the photovoltaic effect, which was firstly described by Alexander-Edmond Becquerel in 1839. As it is reviewed in chapter 3 about Solar Energy Harvesting, the photovoltaic effect has the same quantum nature as the photoelectric effect, so both can only be described by considering that the energy of the electromagnetic radiation is quantized in quanta called photons, with an energy hν, as it has been explained before (equation 2). As it is shown in figure 3.a, photovoltaic effect takes place at the core of the cell, which is found at the junction of the two semiconductors that integrates a typical PV cell. When an individual photon interact with an individual electron at the valence band of the semiconductor, the energy of the photon (and the photon itself) can be absorbed by the electron to get promoted to the conduction band, leaving a hole in the valence band. This process called, photo-generation of an electron-hole pair, is only possible if the photon energy is at least equal to the energy of the band-gap (energy distance between the conduction and valence band). The population of photogenerated electrons and holes is then driven by the electrical field in the depletion zone of the PN junction and can eventually contribute to a photovoltage and the corresponding photocurrent, when an electric load is connected to the PV cell. In this case, both the photovoltage and the photocurrent are *dc* magnitudes and their product gives directly the electrical power converted by the PV cell.

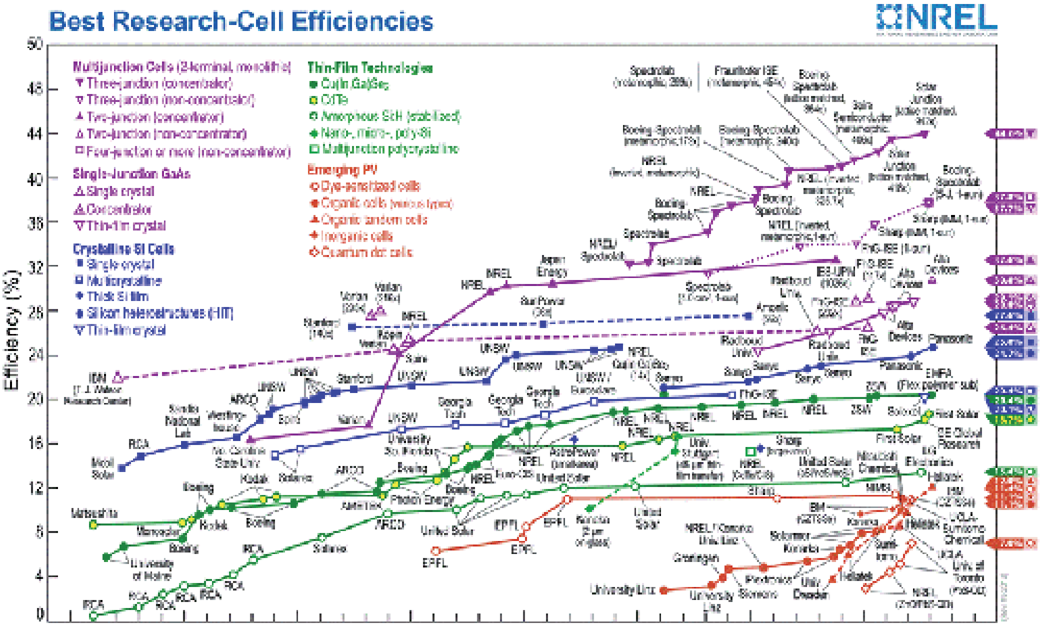


Figure 2. Efficiency evolution of the main photovoltaic technologies (from *National Renewable Energy Laboratory*).

Instead, the radiofrequency rectenna (RFR) technology is based on the combined operation of two basic elements: an electrical rectifier that follows an electromagnetic antenna (rectenna). The operation principle (figure 3.b) does not require quantum mechanics to be explained since, in this case, electrons in the metallic antenna are already in the conduction band, and do not need to be promoted in energy by absorbing photons from the electromagnetic radiation. In this case, the phenomenology is better explained by the interaction between the electrons in the antenna and the electric field of an electromagnetic incident wave. Similarly to PV technology, in rectenna technology matching conditions must be also satisfied. Now, the wavelength of the EM incident wave has to be a multiple of the antenna characteristic length in order to induce a resonant electrical current in the antenna. As opposite to a PV cell, an antenna will generate at its output both an *ac* voltage and an *ac* current. For this reason, a rectifier is needed as the first basic electrical component to transform *ac* values into *dc* values.

Optical rectenna (OR) technology can be considered as a particular case of rectenna technology where the frequency of the electromagnetic radiation involved is in the optical range. So, from this point of view, RFR technology covers the radiofrequency part of the electromagnetic spectrum and OR the optical part (figure 4). However, as it will be described in a next section of this chapter, OR technology cannot be considered just as an extrapolation of the RF rectenna concept to the optical range, since neither the antenna element, in this case a nanoantenna, nor the rectifier, typically a metal-insulator-metal (MIM) diode, have exactly the same properties of the RF counterparts. New physics such as plasmon resonances have to be taken into account in the optical antenna (OA), an antenna with characteristic lengths in the nanometer range (nanoantenna) to match the wavelengths of light radiation. Also special structures and materials are needed to achieve response times short enough to rectify signals in the THz range, which are induced in the nanoantenna element by the incident optical radiation.

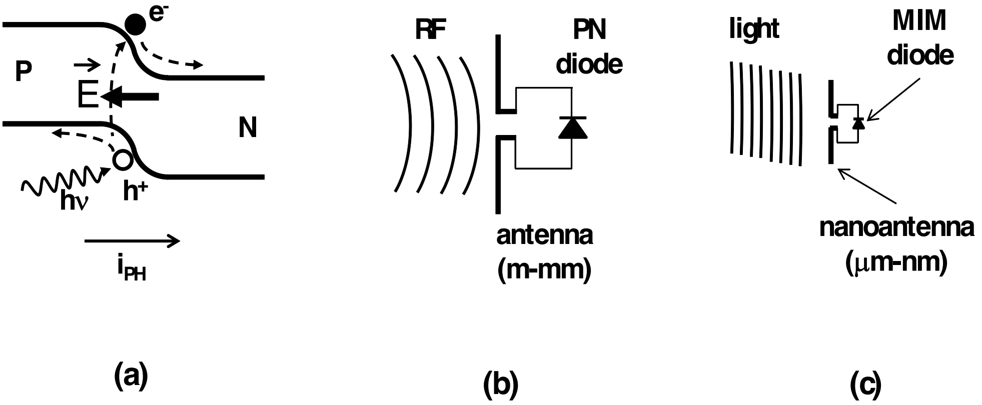


Figure 3. Scheme of the operation principle of three different technologies used to harvest energy from the electromagnetic spectrum. (a) Photovoltaic technology (PV), (b) radiofrequency rectenna technology (RFR) and (c) optical rectenna technology (OR).

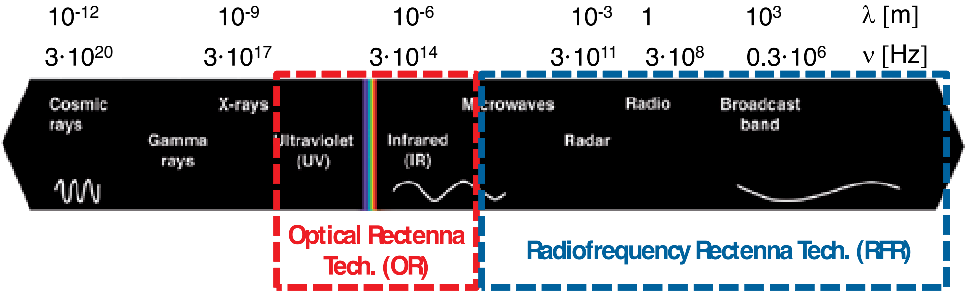


Figure 4. Spectral range coverage of the RF and optical rectenna technologies.

When used as light detectors, optical antennas involving rectifiers perform quite well in several specifications, especially in those related with their intrinsic electromagnetic nature. Table 6 shows these figures for a few technologies working in the visible and the infrared. We may already see in this table that the responsivity of optical antennas is lower than the rest of technologies. This figure is in accordance with the low efficiency of rectennas that has been observed in actual experiments involving MIM, or Metal-Insulator-Insulator-Metal (MIIM), junctions as transducers. Summarizing this table we may say that optical antennas are point detectors, very fast, work at room temperature, can be integrated with some other elements and devices (for example with focusing optics), and they present a broad tuneability, and a remarkable selectivity in direction and polarization.

	Visible CCD/ CMOS	**MIM Junctions**	**Avalanche Photodetectors**	**Pyroelectric Detectors**	**Bolometric Detectors**	**Optical Antennas**
Size	$10^2\lambda^2$	$10^2\lambda^2$	$10^2\lambda^2$	10^1-$10^2\lambda^2$	10^1-$10^2\lambda^2$	10^{-2}-$10^0\lambda^2$
Polarization selective, directivity, tuneability	No	No	No	No	No	Yes
Cooling	Better performance	No	Better performance	No	No	No
Responsivity	10^3-10^4 V/W	0,7-0,9 A/W	0.7 A/W	10^3-10^4 V/W	10^3-10^4 V/W	0.1 V/W
Time response	100 ns	10 ps	9 ps	400 µs	400 µs	1 ps

Table 6. Four different photodetection mechanisms are compared with optical antennas technology.

Nowadays, space in urban areas, including work and home environments is strongly packed with EM radioelectric waves at various bands and spectral regions: besides the ubiquitous presence of radio and TV bands, cell phones and personal communications devices, a myriad of wi-fi stations, Bluetooth gadgets, and remote emitters and detectors produce a non-negligible amount of EM energy flowing around us. Then, from a harvesting point of view this energy could be recycled and properly used by electronic systems with ultra-low power requirements. This strategy may work in those environments with strong RF signals, where signal-to-noise ratio of other operative elements is not compromised. This idea of RF and microwave recycling has been developed some time ago in the form of antenna arrays and half or full wave rectifiers. Optical rectennas can be seen as an evolution and transposition of those designs and devices already working in the microwave region. In this band some designs have demonstrated more than 75% of efficiency when used for power transmission [8]. These figures are reduced when considering broad-band antennas designed to recycle microwave energy from ambient background.

Unfortunately, so far the efficiency number obtained at those frequencies have not been replicated at infrared or visible frequencies. The reasons are mostly derived from the inherent behaviour of materials when frequency increases. Besides, the difficulties of designing THz electronics and oscillators, metals begin to behave as dispersive materials and the currents built on their surface penetrates within the structure.

In order to place the reader in a position to make an educated guess on the different technologies we present here a brief comparison among the photovoltaics, radiofrequency rectifiers, and optical rectennas (optical antennas coupled to rectifiers)

Photovoltaics: Direct conversion of light into electric power using the photovoltaic effect exhibited by semiconductor materials.

- Efficiency: The theoretical limit is around 41% for single junction solar cells, and reaches 87% for multiple junctions.
- Pros: Well established and mature technology. Fabrication issues have been solved due to the intrinsic relation with semiconductor technology.
- Cons: The performance is strongly dependent on temperature, especially for multiple junction cells.

Radiofrequency Rectifiers: Direct conversion of light into electricity using a rectifier working at radio or microwave frequencies.

- Efficiency: The limit is set around 85%. Practical devices have been demonstrated with an efficiency larger than 75%.
- Pros: Well known basic mechanism of rectification. Fabrication can be made using standard photolithography on dielectric substrates.
- Cons: Polarization and spectral selectivity.

Optical Rectennas: Direct conversion of light into electricity using rectifiers working at optical frequencies.

- Efficiency: The theoretical limit is around 85%.
- Pros: Antenna theory and its scaling to optical frequencies is known and antenna-coupled detectors have been demonstrated in the infrared and the visible. Minimum size of about λ^2, allowing very high packaging. No dependence with temperature. Metals are used for fabrication with some advances in the use of conducting graphene.
- Cons: The efficiency of working devices is well below 85%. Barrier rectifiers are not able to follow optical frequencies and behave as square law rectifiers. Further advances are needed to have feasible rectifying mechanisms. Nano-fabrication technologies are necessary (nano-imprint could solve large scale fabrication numbers).

1.3. Historical overview of wireless power transmission

The precedents of the rectenna technology are found in the first attempts to transmit power through radio waves. The early history of RF power transmission dates from the experiments of Heinrich Hertz (1857-1894). Hertz was the first to rigorously prove the existence of electromagnetic waves. The experiments carried out for this purpose (figure 5) were based on transmitters and receivers of radio pulses that were combined with reflectors to create standing waves between the emitter and the receiver. In such experiments, *dc* power was converted to UHF radio waves by means of an LC oscillator connected to a device called *spark gap*. The emitted UHF EM wave from the dipole antenna was directed to the receiving antenna, an open loop ring with also a spark gap, by means of a parabolic reflector. When the emitted wave impinged on the receiving loop, a current was induced and a spark was produced. Hertz was in this way able to verify experimentally the existence and propagation in free space of EM radiated waves and to measure that such propagation is produced at the velocity of light.

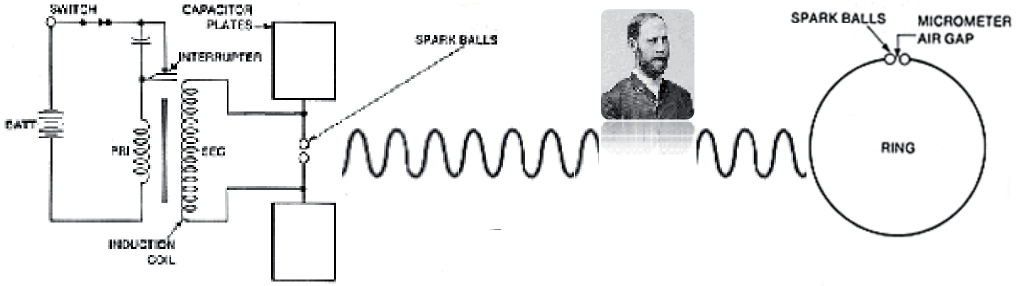


Figure 5. Scheme of the experiments carried out in 1887 by Heinrich Hertz to demonstrate the existence and propagation of electromagnetic waves in free space.

It was some years later, at the turn of the century, that Nikola Tesla (1856-1943) became interested in transmitting electrical power from one point to another wirelessly. Several famous attempts are described elsewhere. In the first one, which was carried out in 1899 at the Colorado Springs Laboratory, an approximately 60 m mast antenna with a 1 m diameter copper ball on top was built. An enormous coil that was fed with 300kW of electrical power got in resonance at a frequency of 150 kHz. When this coil was connected to the mast antenna, an RF potential of 100 MV with respect to the Earth was produced. The only record from this attempt were the discharges from the sphere to ground (figure 6 left), but no data about the power radiated and the power collected at a certain point were reported.

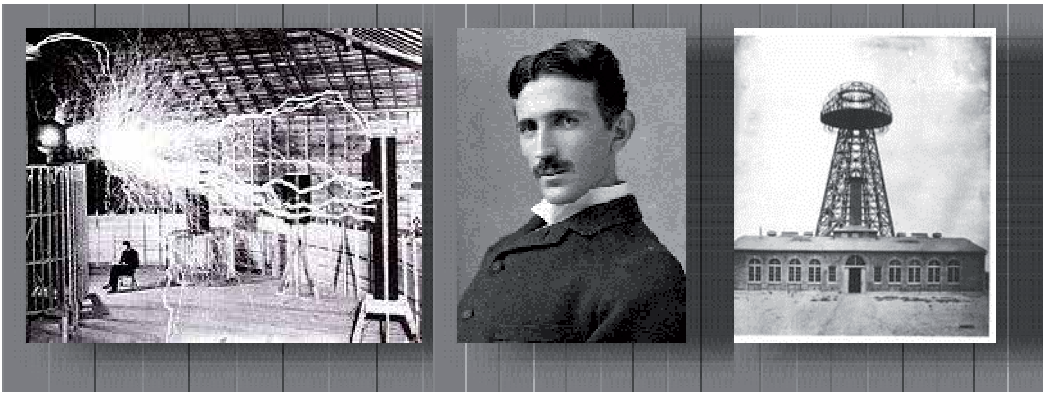


Figure 6. Pictures of the Colorado Spring Laboratory experiments (left), Nikola Tesla (center) and Wardenclyffe plant (right).

Starting in 1901, a similar frustrated attempt to transmit power wirelessly was performed by Tesla in the Wardenclyffe plant at Long Island (figure 6 right), New York. In this occasion, a wooden tower 46 m tall was built to place a 30 m in diameter doughnut-like copper electrode. With this giant installation, Tesla wanted to transmit electrical power across the Atlantic, from the USA east coast to Great Britain. In 1914 the tower was demolished after Tesla lost funding to continue this project.

During the 1930's decade, less ambitious and more controlled experiments performed in the Westinghouse Laboratory led H.V. Noble to successfully transfer hundred watts of power between two 100 MHz dipoles placed 1.5m apart.

A retrospective analysis shows that the initial Teslas's failures and the lack of a clear demonstration of wireless power transfer during the first half of the past century is because power transfer starts to be efficient at the microwave range and above in frequency. At that time the technology to generate power in this wavelength range was not developed enough. It was not until the development of the klystron and the magnetron, combined with the end of the World War II that power transfer technology could start to be notably unfolded. A detailed description of this ramp-up period of the modern history of wireless power transfer is done by one of the most prominent protagonists, William C. Brown [9]. Brown is famous by the invention of the crossed-field amplifier, also known as *Amplitron*, but he can be also considered as the pioneer of the microwave power transmission and the first in developing a rectenna. Most of Brown's achievements were carried out at the Raytheon Company and at the Jet Propulsion Laboratory (NASA), and were mainly driven by two applications respectively: the Raytheon Airborne Microwave Platform (RAMP), a microwave-powered helicopter and the solar-power satellite (SPS), with microwave power transfer to the Earth. The requirements demanded by both applications gave rise to the development of the Amplitron and the first rectenna, as solutions to the generation of high Continuous-Wave (CW) powers of microwaves to be transmitted and to the direct conversion of the received microwaves into *dc* power, in order to drive the motors of the helicopter rotor blades. Thus, in 1964, the first flight of a helicopter prototype was demonstrated. It was propelled by the 270W *dc* provided by a 1.4 kg array-like rectenna integrated by 4480 1N82G semiconductor diodes in a 0.4 m^2 area, which corresponded to a power to mass ratio of 5 kg/kW. Eight years later, in 1968, the improvements introduced by the use of Schottky diodes produced an enhancement of this ratio of one order of magnitude. Finally, in 1983, the introduction of the thin film etched-circuit rectenna technology made possible *dc* to *dc* efficiencies of 85% and power to mass ratios of 1kW/kg.

Further advances in the wireless power transfer technology have been made during the last part of the past century and the beginning of the present one. As a consequence of this last evolution, a company called WiTricity Corp. and a technology called passive Radio Frequency IDentification (RFID) have become the two most meaningful examples of successful application of wireless power transfer.

WiTricity Corp. was born in 2007 to commercialize the applications of a technology developed at Massachussetts Institute of Technology (MIT) by Professor Marin Soljačić and co-workers [10]. The operation principle of this technology is based in the non-radiative power transfer between two self-resonant coils operating in the strong coupling regime.

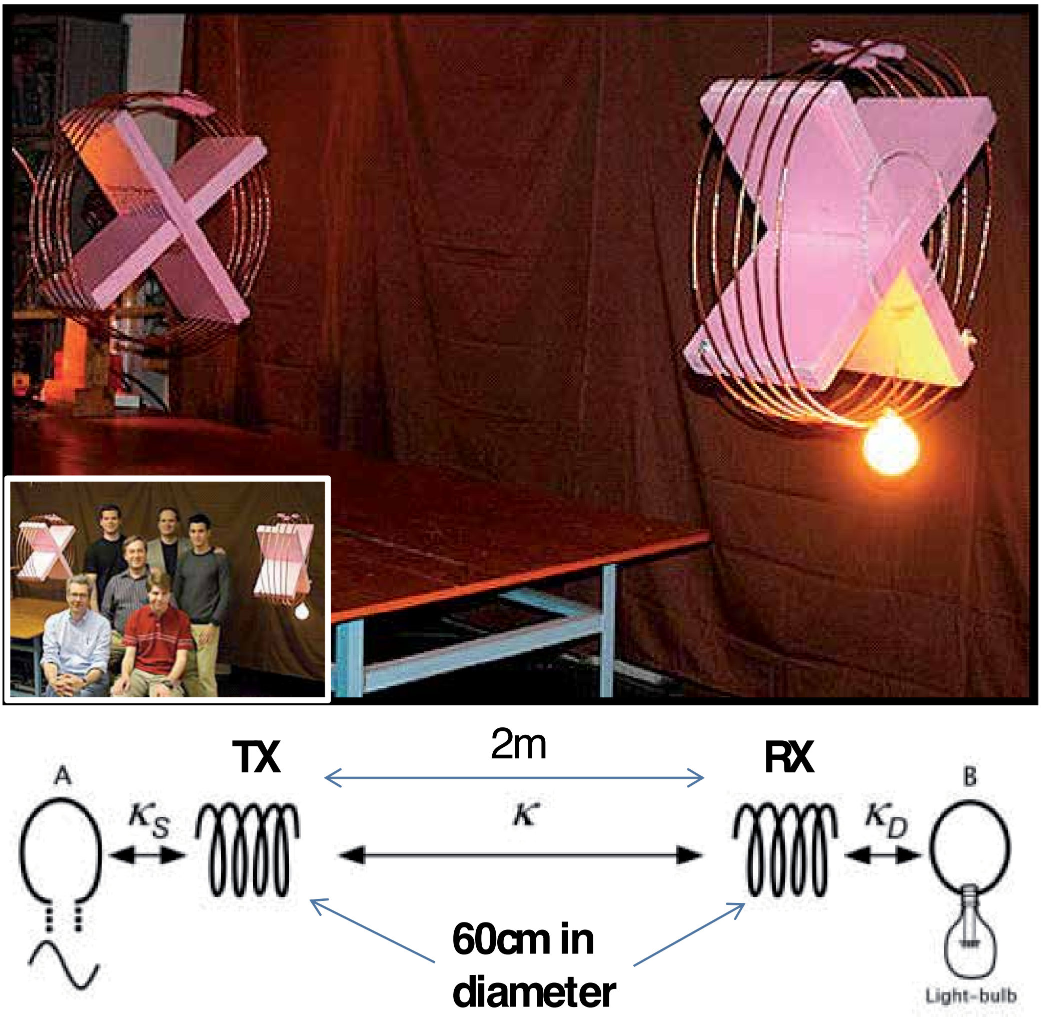


Figure 7. Picture (top) and scheme (bottom) of the WiTricity concept experimental setup. In the inset, the team of Prof. Soljačić at MIT is placed between both coils during operation, trying to demonstrate that the technology is harmless.

In figure 7, the setup used to demonstrate the WiTricity concept is shown. A single loop, A, connected to a sinusoidal signal generator is magnetically coupled to a secondary 5-turns emitter copper coil 60 cm in diameter. An identical receiver coil is coaxially placed at a 2m difference from the emitter, and also coupled capacitivelly to a secondary single loop connected to a 60W bulb load. The system, which is designed to resonate at 9.9 MHz, transfers the 60W of power needed to light the bulb with an efficiency of around 45%. At a shorter distance of 3 ft, the 60W are transferred with an efficiency of 90%.

Finally, RFID technology is also drawing on wireless power transfer technology [11]. In the active RFID variant, the RFID transponder, also called "tag", get the energy from a battery to supply its Application-Specific Integrated Circuit (ASIC). By contrast, in a passive RFID technology, the voltage generated in the tag antenna by the transmitted RF signal during the periods of unmodulated carrier is converted to a dc voltage. This voltage is used to power up the active ASIC chip circuitry which controls the input impedance of its front end. Communication between the base station (RFID reader) and the active tag is based on the modulation of the back-scattered signal produced by the toggle of the input front end impedance between two states (figure 8).

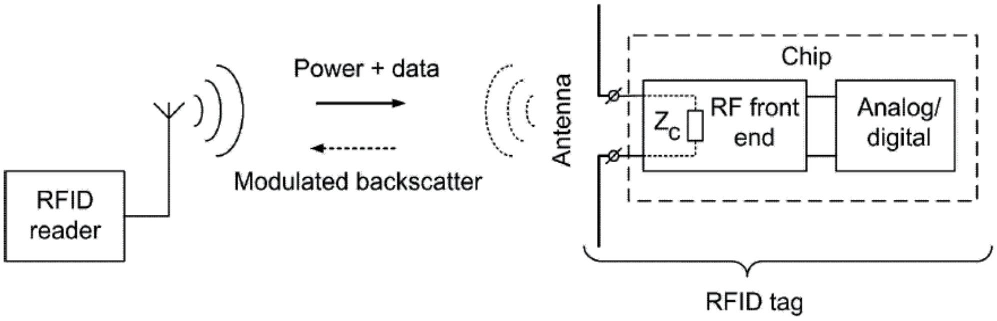


Figure 8. Scheme of an RFID system. When RFID is passive, the power transmitted by the reader during the unmodulated periods is converted in a dc power at the input of a passive circuitry of the tag to supply the rest of the active chip [11].

2. Electromagnetic radiation energy harvesting. The rectenna approach

As it has been pointed out above, a rectenna is the basic element of the RF and optical rectenna technologies. It basically consists (figure 9) of an antenna, in charge of efficiently collecting the energy emitted from a radiative source in the EM spectrum, and a diode, in charge of rectifying the *ac* voltage induced at the antenna terminals by the EM radiation. Eventually, a low pass filter follows the diode in order to obtain a *dc* voltage from the rectified signal. Usually, a *dc-dc* converter is also needed to adapt the voltage levels of the filter output with the level required by the application, represented in figure 9 by its equivalent load. As in most of the energy harvesters, control electronics will manage the flow of energy from the *dc-dc* converter to the application load or to a storage device, usually a battery, depending on whether the energy harvested by the rectenna can satisfy the application demand or, instead, it is better to store the harvested energy until the load demand could be satisfied.

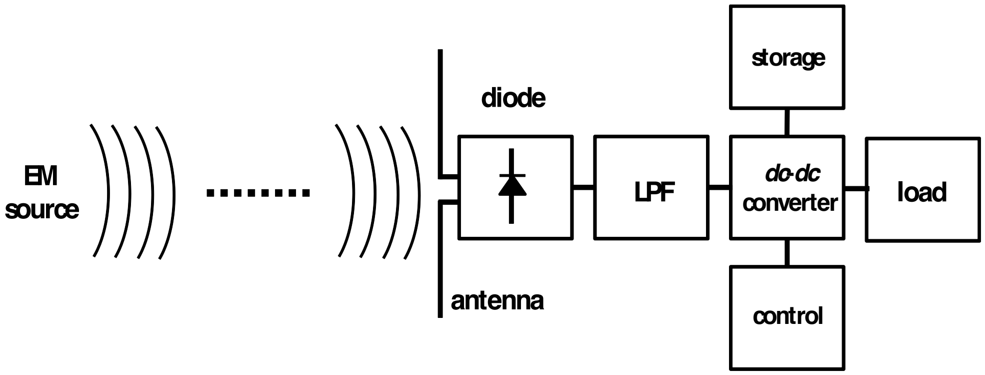


Figure 9. Block diagram of a rectenna.

The core of the rectenna, i.e. the antenna and the rectifier, can be replicated in an array configuration in order to improve the collection efficiency. Thus, a 2D array of identical rectenna elements can be connected in series or in parallel to increase the effective collection area and to increase the output voltage or the output current respectively [12]. Another 2D array configuration strategy is based on combining rectenna elements with different characteristics in order to match the different wavelengths of an EM sources set.

2.1. The antenna as transducer element

An antenna is a device made to transmit and/or receive EM waves. By converting an electric current into an electromagnetic field on one end and converting this EM field into a voltage on the other, a pair of antennas gives the capability of making a wireless link between two points.

The most important parameters of a transmitter antenna are described below:

- Impedance: given the fact that the antenna must be connected to a transmitter, through a transmission line, and radiate the maximum amount of power with the lowest losses, the impedance of the transmitter, the transmission line and the antenna must be the same. The antenna itself introduces losses in the system, normally ohmic. In almost all the EM antennas the input impedance can be computed as the sum of the losses (R_Ω) and the radiation (R_r) resistance

$$Z_{in} = R_r + R_\Omega \tag{8}$$

The total delivered power to the antenna should then be calculated as:

$$P_T = \mathrm{P_r} + P_\Omega = I_{in} \cdot R_r + I_{in} \cdot R_\Omega \tag{9}$$

- Efficiency: given the losses and radiation resistances the efficiency of an antenna can be computed as

$$\eta_e = \frac{R_r}{R_r + R_\Omega} \tag{10}$$

- Directivity: the directivity of an antenna is defined as the relation between the radiated power density in one particular direction and the radiated power density which would radiate an isotropic antenna emitting the same power.
- Radiation pattern: the radiation pattern of an antenna describes the relative field strength of the radiated EM waves in all the directions from the antenna, at a fixed distance. For directional antennas the radiation pattern shows that there is a particular direction on which the antenna emits more efficiently. For omnidirectional antennas the radiation patter is almost equal for all directions.
- Gain: compared to an isotropic radiator, which will equally distribute the radiated power in all directions, real antennas with either a directional or omnidirectional radiation pattern will radiate less power in some directions and more in others. Therefore it can be considered that there is a gain between the different radiation directions. This gain can be defined as the ratio between the transmitted signal strength value at the more efficient direction and the value using a reference antenna. If the reference antenna is an isotopic source the used units will be dBi.
- Polarization: the polarization of an antenna is defined as the orientation plane in which the radiated or absorbed electric field vibrates with respect to a reference plane, for example the Earth's surface. It is determined by the physical construction of the antenna and its orientation, especially by its radiating element. The most common polarizations are linear and circular, which are particular cases of elliptical polarization. In the first one, the electric field vector stays in the same plane whereas for the second one it appears to be rotating with a circular motion around the direction of propagation.
- Bandwidth: most of the EM antennas operate efficiently over a relatively narrow frequency span due to their geometry. As a consequence, they must be tuned in order to have the same frequency band operation as the electronic system at which they are connected.

The most important parameter of a receiver antenna is

- The effective area: an antenna extracts power from the wavefront of an EM wave, thus it represents a certain capture area or effective area. This area is defined as the relation between the power delivered to the receiver circuitry and the power density of the incident wave.

Additionally, antennas comply with the law of reciprocity. This law states that given two identical antennas placed at some distance, each of them can be operated either as a transmitting antenna or as a receiving antenna. Suppose that the one working as a receiver is kept intact, while the performance is modified so that, for a fixed amount of radiated power, the signal received by the other antenna changes by a factor. If the same modified antenna is used for receiving the transmitted signal by the unmodified one, its performance will also be changed by the same factor. This theorem can be formally derived from Maxwell's equations and its validity can be easily verified.

As a consequence of this law, all the previous described antenna characteristics (efficiency, radiation pattern, gain, polarization, bandwidth and effective area) are the same whether the antenna takes part in a transmitter or a receiver scheme. Besides, when designing optical antennas, this reciprocity law is used to simplify the calculation, for example, when calculating the radiation/receiving patterns and some other important parameters of the antenna.

The simplest designs of radiofrequency antennas are developed as wires or loops of conductors properly arranged and connected to an electronic element to produce and detect electromagnetic radiation with a wavelength scaled to the size of the antenna. An important step forward was made when planar structures used the resonant properties of metal patches and planar strips. These could be fabricated using photolithography in a very similar manner as it was done with printed circuits. When the frequency increases the resolution of manufacturing techniques also increases to produce thinner and finer structures. Then, when moving to terahertzs and infrared frequencies only nanofabrication techniques are suitable to realize those antennas. The reasons are twofold. On the one hand is the shrinking of the wavelength towards the nanometric scale, and on the other hand the quality of the finishing elements in terms of roughness and surface smoothness, which may interact with the currents and scatter the charge carriers around non useful directions. These nanophotonic devices are still considered as antennas because they can produce or detect electromagnetic radiation using wires or resonant patches. In a simple manner, optical antennas are defined as resonant structures able to produce an electric signal related with the incident optical radiation.

At the same time, taking into account the re-emission of electromagnetic radiation by resonant structures, it is possible to define a new kind of element that changes the properties of the light that interacts with it. We will name these elements as belonging to the "resonant optics" area. Conventional, reflective or refractive, and diffractive optics relies on the geometrical and wave models of light. Then, resonant optics uses the electromagnetic interaction of light waves with geometrical structures, typically fabricated with conductors, which work building currents up. They have been used as frequency selective surfaces, polarization elements, or phase shaping devices. Some of the designs are scaled versions of their microwave counterparts, where they were first demonstrated. In the recent past we have seen a growing number of scientific contributions where these elements are analyzed and exploited for a variety of applications. At the same time, they have been denoted with different names, mainly depending on the origin of the research teams that develop them: metamaterial surfaces, flat optics, 2.5D photonic crystals, etc.

Both optical antennas and resonant structures are determined by geometry and material parameters. Geometry mostly drives the polarization and spectral selectivity. Intrinsically, the size of the antenna is related with the wavelength. Therefore, when considering infrared and optical radiation, optical antennas become, by nature, nanophotonic devices. For example, a dipole antenna has a length of a few microns for far infrared radiation, and a few hundreds of nanometers for visible light. At the same time, the width of the dipole is limited by fabrication constrains and can be as narrow as a few tens of nanometers. Consequently, the area of detection of the incoming radiation, which extends a little farther from the antenna itself, is about l^2, depending on the geometry of the antenna. The far-field pattern of optical antennas resembles that of their radiofrequency counterparts.

Also, as it happens with planar antennas written on dielectric substrates, the responsivity of optical antennas is larger when light is incident from the substrate side than when it is incident from the air, mostly because of the larger electric permittivity of the substrate. The ratio between the power radiated, or received, by an antenna located between two media of electric permitivities ε_1 and ε_2 is [13,14]

$$\Gamma = \frac{P_1}{P_2} = \left(\frac{\varepsilon_1}{\varepsilon_2}\right)^{3/2} \tag{11}$$

Besides, the effective wavelength at which the resonance takes place moves because of this situation. This effective wavelength is given classically as

$$\lambda_{eff} = \frac{\lambda_0}{\sqrt{\frac{\varepsilon_1 + \varepsilon_2}{2}}} \tag{12}$$

and when considering the plasmon resonances as

$$\lambda_{eff} = n_1 + n_2 \frac{\lambda_0}{\lambda_p} \tag{13}$$

where n_1 and n_2 depend on the geometry and material parameters, and λp is the wavelength of the plasmon resonance [15]. This equation already shows the influence of the material parameters in the performance of optical antennas. At the infrared and visible frequencies metals are no longer perfect conductors and behave as dispersive materials [16]. This means that the radiation losses increase and the surface currents penetrate deeper within the materials. However, metals also present interesting phenomena at optical frequencies. The optical radiation can excite the collective resonance of the charge carriers. They now, oscillate as a unique particle that is named as plasmon. The plasmonic resonances have been devoted increased attention along the past years, producing a myriad of papers and novel applications. In the area of optical antennas, plasmons play an important role because of their occurrence at optical frequencies.

Although in this chapter we focus our attention on optical antennas, we have to mention the important role that resonant structures may play in the improvement of photovoltaic solar cells. It is known that when populating a surface with metal nano-structures, the interaction of light with the structure changes. If the surface is that one of a solar cell, photons can be scattered by the resonant structures and the optical path within the material is enlarged. Also the enhancement of the electric field near the nano-structures may increase absorption, or directly, those photoexcited electrons can be injected into the cell, contributing to the total current delivered by the cell [17]. Some advances have been made reporting 50% increase of the transmittance of the surface using plasmonic nanoparticles [18]. This path is quite promising for improving the performance of traditional photovoltaic cells.

2.2. Rectifying devices and technologies

Rectification is commonly performed by p-n junction diodes when RF radiation is in the kHz-MHz low frequency range. However, when operation frequencies are in the GHz-THz range, semiconductors and devices with shorter transit times and lower intrinsic capacitances like GaAs Schottky diodes are needed. Typical maximum operation frequency of Schottky diodes is 5 THz and although theoretical efficiencies approach 90%, only values of just around 50% have been demonstrated experimentally.

In most common rectifying situations, which correspond to low frequency and high power conditions (LFHP), diodes produce a half-wave rectification with an efficiency given by

$$\eta_{LFHP} = \frac{1}{1 + \frac{v_D}{2v_{dc}}} \tag{14}$$

where v_D and v_{dc} are respectively the voltage drop across the diode and the output rectified *dc* voltage.

However, in the opposite conditions of high frequency and low power (HFLP), which corresponds to those of RF signals to be harvested by a rectenna, rectification is much more complex Since: first the incident power is not only low but fluctuant in value and, second, because matching of the antenna and the diode is an issue. Consequently, the power at the output of the rectenna, P_{dc}, will be obtained from the incident RF power, P_{RF}, by

$$P_{dc} = P_{RF} \cdot \eta_{HFLP}(P_{RF}, \rho) \tag{15}$$

where the rectification efficiency, η_{HFLP}, depends explicitly on P_{RF} and on the antenna to diode matching ρ. This explicit dependence cannot be described analytically with a closed expression like in the LFHP case and the problem of P_{dc} prediction has to be solved by simulation. Time-domain analysis has been successfully applied to address this problem for single frequency or narrow-band rectenna applications [19]. However, a frequency-domain approach based on the harmonic balance (HB) method is more appropriated for wide-band applications [12], where some characteristics of the diode as the nonlinearity of its capacitance, the reflected harmonic energy at input/output or the self-biasing effects start to be relevant.

In the case of optical antennas, they work by combining the effect of two physical mechanisms. One of them is the coupling of the optical radiation to the device. This task is in charge of the metal structure. The other is the transduction mechanism used to provide the output signal. So far, two main types of transducers have been demonstrated. The bolometric response of the material is used to produce a change in the voltage measured from the device. However, this mechanism is dissipative and it does not provide a positive balance of power, and therefore it is not useful for harvesting applications, at least directly. As an interesting outcome of the Joule dissipation we find that when incorporating resonant elements to thermoelectric pairs, the combination of localized heating and the Seebeck effect can be of use to produce electric power [20,21]. The other mechanism is the rectification of currents using a diode. This is the case implemented in rectennas. Metal-Insulator-Metal (MIM), and Metal-Insulator-Insulator-Metal (MIIM) have been demonstrated as effective diodes from the microwave to the visible wave ranges [22, 23, 24, 25]. Typical MIM materials are Cr/CrO$_x$/Au, Nb/NbO$_x$/Nb or Al/AlOx/Pt, and state of the art MIM diodes can operate at frequencies up to 150 THz. These junctions work as square-law rectifiers. The rectified current is given as

$$I_{DC} = \gamma \frac{|V_{diode}|^2}{4R_{diode}} \tag{16}$$

where γ represents the non-linearity of the current-voltage curve of the device. Non-linearity, γ, which is defined from the i-v diode characteristic as $\gamma \equiv (d^2i/dv^2)/(di/dv)$, should be at least 3 or larger to start getting reasonable values of conversion efficiency. Figure 10 shows the line transmission schematics of the antenna-diode element. When optimizing this structure for maximum efficiency the impedance of the antenna has to compensate the impedance of the load (the diode). This means that the antenna has to present an imaginary part of the impedance that is not typically included when maximizing the performance.

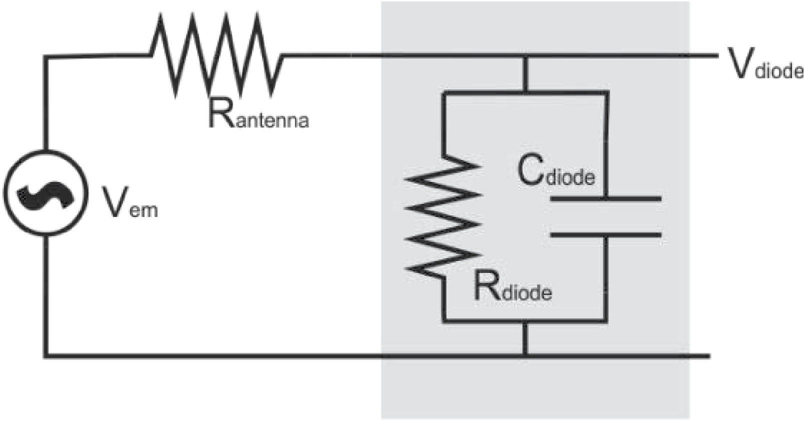


Figure 10. Transmission line schematics of a diode coupled to an antenna. The antenna is considered as having an impedance of R$_{antenna}$.

On the other hand, the cut-off frequency of the element is given as

$$f_{cutoff} = \frac{1}{2\pi RC} \tag{17}$$

where the RC constant should be smaller than 10^{-14} s^{-1} to reach the infrared and optical regions. The junction itself works as a capacitor. To reduce its capacitance we cannot increase the thickness of the junction because in that case tunnelling would be not possible. Because of that, it has a thickness of about 2-3 nm, depending on the insulator. Therefore, in order to have a low capacitance the junction area should be small. On the other hand, to efficiently transfer the power to the load, the impedance mismatch should be corrected. These issues are being addressed in several ways and some promising results have been already published [26,27]. However, when combining MIM diodes with antennas, the conversion efficiency given as the ratio between the power obtained after rectification and the incident power is quite low for these square law rectifiers [28]. The values for this efficiency is lower than 10^{-6}, showing that some better rectifying technologies need to be developed before practical devices based on direct rectification of light, become competitive against other solar energy harvesters. In Figure 11 we show the response of an optical antenna placed in front of a black body at 1000°K, demonstrating the existence of an output signal for this extreme condition.

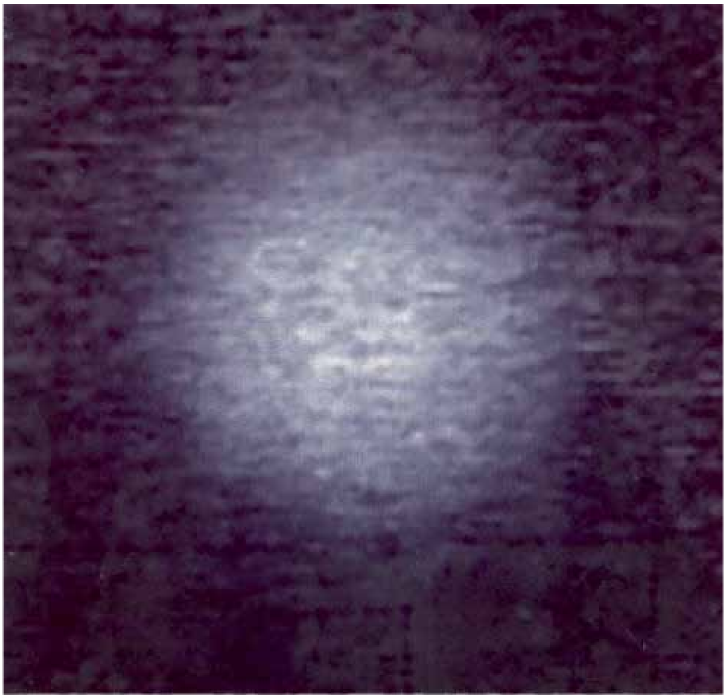


Figure 11. Response of an optical antenna located on the image of a blackbody radiator at 1000°K. This image was obtained by a relay optical system fabricated in ZnSe and having an F/# equal to 1 [29].

A modification of the traditional MIM diode, the travelling-wave metal-insulator-metal diode (TW-MIM), allows obtaining a quantum efficiency of 3.6% in the IR region [30]. The TW-MIM is based on the rectification of the surface plasmon excited by the antenna on a plasmonic waveguide.

Some interesting advances in rectification at terahertz and quasi-optical frequencies have been proposed using a novel approach. The geometric rectifier (see Figure 12) has been demonstrated at GHz frequencies [31]. Here the rectification is given by deflecting the trajectories followed by the charge carriers through an asymmetric channel. The simplest case is an arrow shaped element that selectively directs charge carriers in a given direction, preventing the movement of those carriers towards the opposite direction [32]. These geometric rectifiers need materials and conditions where the carriers exhibit a free mean path longer, or much longer, than the size of the rectifying structure, i.e., a few hundreds of nanometers. Besides the electric properties of metals, exhibiting free mean paths in the range of a few tens of nanometers (for example for Au this parameter is around 20 nm), the conduction properties of graphene can be tailored to produce feasible devices having this effect [33].

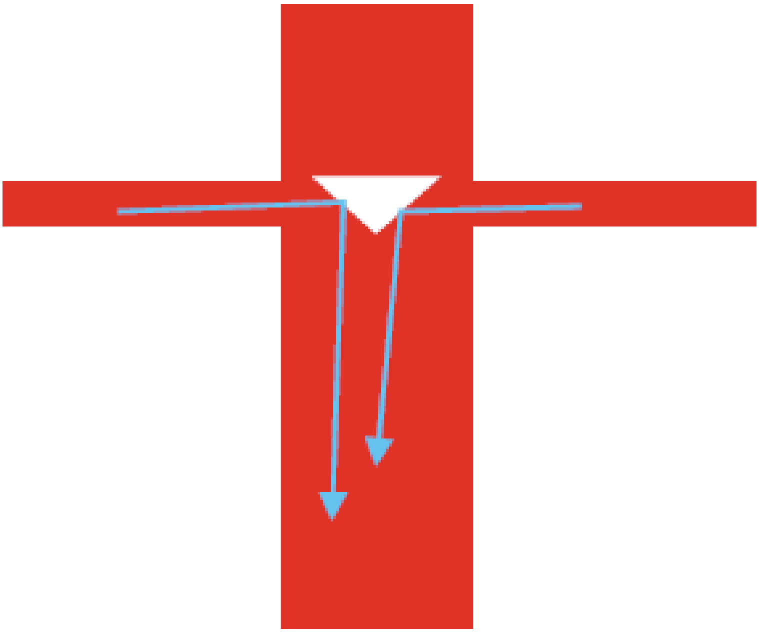


Figure 12. Basic scheme of a geometric rectifier. The horizontal arms of a dipole antenna intersect at the feed point. This feed point is shaped as an asymmetric defect that deflects charge carriers towards the bottom of the geometric rectifier [34].

Also a device based on grapheme, the field effect transistor (G-FET) in common source configuration, can potentially be a promising candidate as rectifier component for optical rectennas. The extremely high mobility of graphene combined with the ambipolar transport properties would allow to implement full-wave rectification at THz in a single device [35].

Finally, two examples of RF rectenna (RFR) and optical rectenna (OR) from the literature are shown in figure 13 in order to compare dimensions and performances.

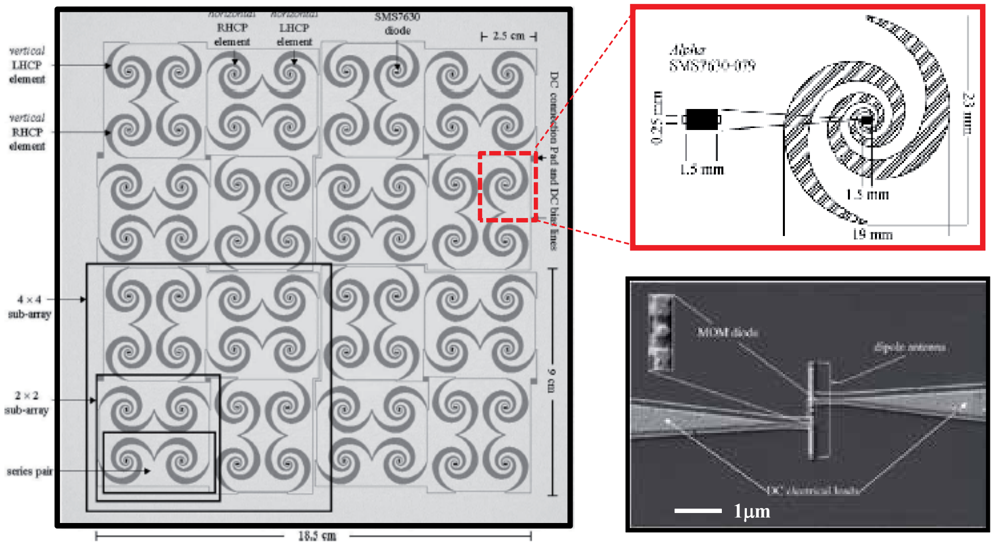


Figure 13. RF rectenna array (left) integrated by 64 single spiral rectennas (right top) [12]. Optical rectenna for the IR range consisting on a dipole coupled to a MOM diode (right bottom) [36].

The RF rectenna of figure 13 corresponds to a rectenna array configuration designed to operate in the 2-18 GHz region and for input power densities from 10 nW/cm^2 to 0.1 mW/cm^2. The spiral rectenna elements are distributed along the 324 cm^2 array area with different orientations in order to harvest energy from different polarized sources. So, considering an effective area of A_{eff}=25 cm^2, then input RF power will vary between P_{RF}=250 nW and P_{RF}=2.5 mW. If rectification efficiencies for such input powers results to be η(250nW)=1% and η(2.5mW)=20%, then the output *dc* power harvested by this array will theoretically vary between 2nW and 450 μW. Other examples of rectenna designs achieved efficiencies from 40-50% operating at 10^{-2} mW/cm^2 [37] to 80% at 10 mW/cm^2 [38].

By contrast, in the optical rectenna (figure 13 right bottom), designed to operate in the IR band, the 19x23 mm^2 area of the RF rectenna element is reduced to less than 1 μm^2. In this case, an Al/AlO$_x$/Pt MOM diode is chosen to implement the rectifier, which is coupled to a dipole 1 μm long nanoantenna. However, as it has pointed out above, the maximum conversion efficiency of 20% shown by the RF rectenna is reduced below 10^{-6} in optical rectennas like this one.

3. Conclusions

In this chapter, the most important concepts needed to understand how energy from the electromagnetic spectrum can be harvested by means of the rectenna technology have been introduced. Main differences with the well stablished photovoltaic approach have been analyzed and a comparative list of pros and cons has been provided. An historical overview of the first works on wireless power transmission has been useful to understand the origin of the rectenna concept. The most relevant technical characteristics of both components of a rectenna, the antenna and the rectifier device, have been also decribed, and the specific features of each element have been explained for the readiofrequency and the optical range.

Acknowledgements

This work has been partially supported by project ENE2009-14340 from the Ministerio de Ciencia e Innovación of Spain and FP7-EU-Project FET-Proactive CA no.270005 (ZEROPOWER).

Author details

Gabriel Abadal[1], Javier Alda[2] and Jordi Agustí[1]

1 Departament d'Enginyeria Electrònica, Escola d'Enginyeria, Universitat Autònoma de Barcelona, Barcelona, Spain

2 Applied Optics Complutense Group. Facultad de Óptica y Optometría, Universidad Complutense de Madrid, Madrid, Spain

References

[1] Shevgaonkar, R. K. *Electromagnetic waves*. Tata McGraw-Hill Education, 2005.

[2] Recommendation, Council. "519/EC of 12 July 1999 on the limitation of exposure of the general public to electromagnetic fields (0 Hz to 300 GHz)."Official Journal L 197 (1999): 1999.

[3] IEEE Std C95. 1-2005. "IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 KHz to 300 GHz." New York: The Institute of Electrical and Electronic Engineers, 2005.

[4] B. Berland, "Photovoltaic Technologies Beyond the Horizon: Optical Rectenna Solar Cell", Report for the National Renewable Energy Laboratory, NREL/SR-520-33263 (2002)

[5] R. Corkish, M. A. Green, T. Puzzer, "Solar energy collection by antennas", Solar Energy, 73, 395-401 (2002).

[6] D. K. Kotter, S. D. Novack, W. D. Slafer, P. J. Pinhero, "Theory and manufacturing Processes of Solar Nanoantenna Electromagnetic Collectors", Journal of Solar Energy Engineering, 132, 011014 (2010).

[7] G. A.E. Vandenbosch, Z. Ma, "Upper bound for the solar enegy harvesting efficiency of nano-antennas", Nano Energy, 1, 494-502 (2012).

[8] B. Strassner and K. Chang, "A circularly polarized rectifying antenna array for wireless microwave power transmission with over 78% efficiency," in *IEEE MTT-S Int. Microwave Symp. Dig.*, (2002), 1535–1538, (2002).

[9] W.C. Brown, *The History of Power Transmission by Radio Waves*. IEEE Transactions on Microwave Theory and Techniques, 32 (9), 1230-1242 (1984).

[10] A. Kurs, A. Karalis, R. Moffatt, J. D. Joannopoulos, P. Fisher, M. Soljacˇic, "Wireless Power Transfer via Strongly Coupled Magnetic Resonances", Science 317, 83-86 (2007).

[11] K. V. Seshagiri Rao, P. V. Nikitin, and S. F. Lam, "Antenna Design for UHF RFID Tags: A Review and a Practical Application", IEEE TRANSACTIONS ON ANTENNAS AND PROPAGATION, 53, 12, 3870-3876 (2005).

[12] J. A. Hagerty, F. B. Helmbrecht, W. H. McCalpin, R. Zane, Z. B. Popovic, "Recycling Ambient Mricowave Energy With Broad-Band Rectenna Arrays", IEEE Transactions on Microwave Theory and Techniques, 52, 1014-1024 (2004).

[13] C.R. Brewitt-Taylor, D.J. Gunton, H.D. Rees, Planar antennas on a dielectric surface, Electron. Lett. 17. 729–730, (1981).

[14] J. Alda, C. Fumeaux, M. Gritz, D. Spencer, G. Boreman, "Responsivity of infrared antenna-coupled microbolometers for air-side and substrate-side illumination", Infrared Physics and Technology, 41, 1-9 (2000).

[15] L. Novotny, "Effective wavelength scaling for optical antennas", Physical Review Letters, 28, 262602 (2007).

[16] J. González, J. Alda, J. Simon, J. Ginn, G. Boreman, "The effect of the metal dispersion on the resonance of antennas at infrared frequencies", Infrared Physics and Technology, 52, 48-51 (2009).

[17] C. Hägglund, M. Zäch, G. Petersson, B. Kasemo, "Electromagnetic coupling of light into a silicon solar cell by nanodisk plasmons", Applied Physics Letters, 92, 053110 (2008).

[18] P. Spinelli, M. Hebbink, R. de Waele, L. Black, F. Lenzmann, A. Polman, "Optical Impedaance Matching Using Couple Plasmonic Nanoparticle Arrays", Nanoletters, 11, 1760-1765 (2011).

[19] T. Yoo and K. Chang, "Theoretical and experimental development of 10 and 35 GHz rectennas," IEEE Trans. Microwave Theory Tech., 40, 1259–1266, (1992).

[20] C. Fu, "Antenna-coupled nanothermopile", M.S. Thesis, University of Central Florida (1998).

[21] G. P. Szakmany, P. Krenz, A. Orlov, G. Bernstein, W. Porod, "Antenna-Coupled Nanowire Thermocouples for Infrared Detectioin"; Proc of the 12th. IEEE International Conference on Nanotechnology (2012).

[22] A. Sanchez, C. F. Davis, Jr., K. C. Liu and A. Javan, "The MOM tunneling diode: theoretical estimate of its performance at microwave and infrared frequencies," Journal of Applied Physics. 49, 5270-5277 (1978).

[23] C. Fumeaux, W. Herrmann, F. Kneubühl and H. Rothuizen, "Nanometer thin-film Ni-NiO-Ni diodes for detection and mixing of 30 THz radiation," Infrared Phys. Technol. 39, 123-183 (1998).

[24] B. J. Eliasson, "Metal-Insulator-Metal Diodes for Solar Energy Conversion," Ph.D. Thesis, University of Colorado (2001).

[25] C. Fumeaux, J. Alda, G. Boreman, "Lithographic antennas at visible frequencies", Optics Letters, 24, 1629-1631 (1999).

[26] Jer-Shing Huang, Thorsten Feichtner, Paolo Biagioni, and Bert Hecht, Impedance Matching and Emission Properties of Nanoantennas in an Optical Nanocircuit" NanoLetters, 9, 1897-1902 (2009).

[27] Yauhen Sachkou, Andrei Andryieuski,, Andrei V. Lavrinenko, "Impedance Conjugate Matching of Plasmonic Nanoantenna in Optical Nanocircuits", Proceedings of the 53rd International Symposium ELMAR-2011, 389-391 (2011).

[28] E. Briones, J. Alda, F. J. Gonzalez, "Conversion efficiency of broad-band rectennas for solar energy harvesting applications", Optics Express, 21(S3), A412-A418, (2013).

[29] J. Alda, "Response of an optical antenna to blackbody radiation", Personal Communication to G. Boreman (April, 1999).

[30] S. Grover, O. Dmitriyeva, M. J. Estes, and G. Moddel, Traveling-Wave *Metal/Insulator/Metal Diodes for Improved Infrared Bandwidth and Efficiency of Antenna-Coupled Rectifiers*, IEEE TRANSACTIONS ON NANOTECHNOLOGY, 9 (6), 716-722, (2010).

[31] A. M. Song, ``Electron ratchet effect in semiconductor devices and artificial materials with broken centrosymmetry", Applied Physics. A, 75, 229-235 (2002).

[32] Garret Moddel, "Geometric diode, applications and method", US Patent Office, Pat. No.: 20110017284 (2011).

[33] Z. Zhu, S. Joshi, S. Grover, and G Moddel, "Graphene Geometric Diodes for Terahertz Rectennas," J. Phys. D: Appl. Phys. in press (2013).

[34] J. Alda, "Geometrical Rectification", Personal Communication to G. Boreman (August, 2006).

[35] H. Wang, D. Nezich, J. Kong, and T. Palacios, *Graphene Frequency Multipliers*, IEEE ELECTRON DEVICE LETTERS, 30 (5), 547-549 (2009).

[36] J. A. Bean, A. Weeks, and G. D. Boreman. "Performance Optimization of Antenna-Coupled Al/AlOx/Pt Tunnel Diode Infrared Detectors", IEEE JOURNAL OF QUANTUM ELECTRONICS, 47, 126-135 (2011).

[37] W. C. Brown, *An experimental low power density rectenna* in IEEE MTT-S Int. Microwave Symp. Dig., 197–200 (1991).

[38] J. O. McSpadden, F. E. Little, M. B. Duke, and A. Ignatiev, *An in-space wireless energy transmission experiment*, in Proc. IECEC Energy Conversion Engineering Conf., 1, 468–473 (1996).

Energy Storage: Battery Materials and Architectures at the Nanoscale

James F. Rohan, Maksudul Hasan, Sanjay Patil,
Declan P. Casey and Tomás Clancy

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57139

1. Introduction

1.1. Energy storage

The intermittent nature of energy harvesting technologies and the low power delivery capability necessitates the integration of energy storage in the overall system design. A means of storing the energy produced in periods of high availability for use in periods of limited harvesting is essential. Not only is storage a major factor in the efficient use of harvested energy but it is also needed on a wide variety of time scales – seconds, minutes, hours, days – reflecting the nature of the intermittency of these sources. Correspondingly, a variety of storage technologies with different storage capabilities and response times are available.

The main storage options appropriate to ICT zero power devices are:

Batteries: Batteries are electrochemical devices using chemical reactions to generate power.

Hydrogen: Requires hydrogen production, compression, storage and power generation through fuel cells

Super-capacitors and Ultra-capacitors: Energy is stored as accumulated charge.

Batteries and hydrogen are capable of operation over the widest range of energy and power densities and thus application areas. This chapter will focus on the battery energy storage options as the most developed option with the potential for further improvements and applications in ICT devices.

Batteries are electrochemical devices that store electrical energy by directly converting it to a chemical form. Examples include lead acid, nickel-cadmium, nickel-metal-hydride, lithium-ion, sodium-sulphur, metal-air and flow batteries. The history of battery commercialisation is represented in figure 1, which shows the energy density per unit volume and weight for the most common rechargeable battery systems. It can be seen that the progression has been to lighter, more energy dense systems using less harmful chemicals and a greater cycling efficiency.

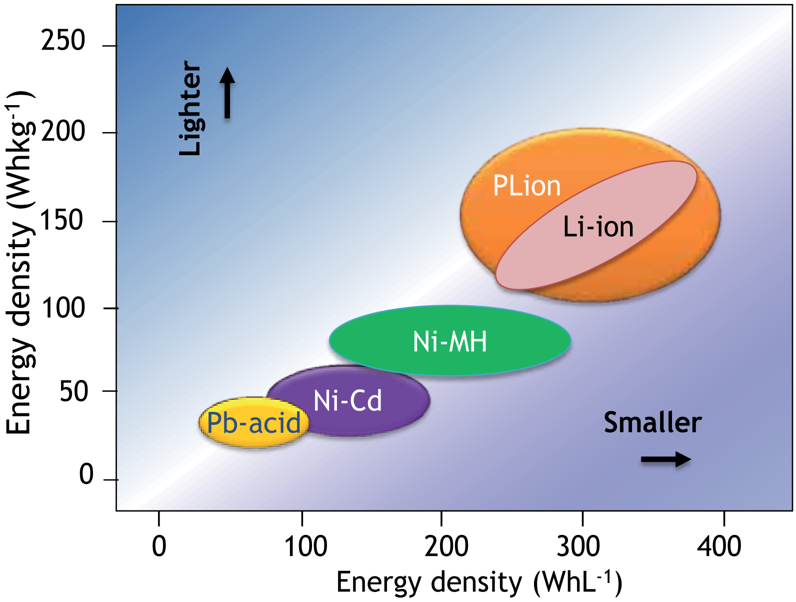


Figure 1. Commercial evolution of rechargeable batteries to higher energy density.

1.2. Theoretical potential, capacity and energy of batteries

The amounts of electric energy per mass or volume that a battery can deliver depends on the chemical energy stored within the electrodes. During discharge a redox reaction occurs, which gives a change in the Gibbs free energy (ΔG) of the system. High energy conversion efficiency from chemical to electrical within the cells is desirable. However, the energy available is lower than the stored chemical energy, due to polarisation of the electrodes. [1] There are mainly two types of polarisations, the activation polarisation required to drive an electrochemical reaction and secondly, the concentration polarisation due to the differences in concentration between the reactants and products at the electrode surface and in the bulk of solution due to mass-transfer. The polarisation decreases the total available energy from the cell, which is lost in the form of heat. The activation and concentration polarisation can be calculated if electrochemical parameters and mass-transfer data are measurable. However, this is practically rather difficult to measure due to the complicated architecture of the electrodes.

The electrodes are normally made of composite active materials with bindera and conductivity and performance improving additives. Another key factor is the internal resistance of the cell which also contributes to the current drain capability and overall performance of the battery. The internal resistance causes a voltage drop during the operation, generally referred to as ohmic polarisation and its magnitude is proportional to the current delivered. Ohm's law applies when a cell with potential E is connected to an external load R and can be expressed as

$$E = E_o - \left[(\eta_{ct})_a + (\eta_c)_a \right] - \left[(\eta_{ct})_c + (\eta_c)_c \right] - iR_i = iR \tag{1}$$

where

E_0= electromotive force or open-circuit potential of the cell (OCP)

$(\eta_{ct})_a$, $(\eta_{ct})_c$ = activation polarisation at the anode and cathode

$(\eta_c)_a$, $(\eta_c)_c$ = concentration polarisation at the anode and cathode

i = load current

R = internal resistance of cell

As can be seen from the Eq. (1) that output potential is lower than the open-circuit potential (OCP) due to the electrode and ohmic polarisation. The electrode and ohmic polarisations are small when the load-current is very low, and in that case the cell may operate close to OCP and deliver most of the total theoretical energy as electric energy, see Fig. 2.

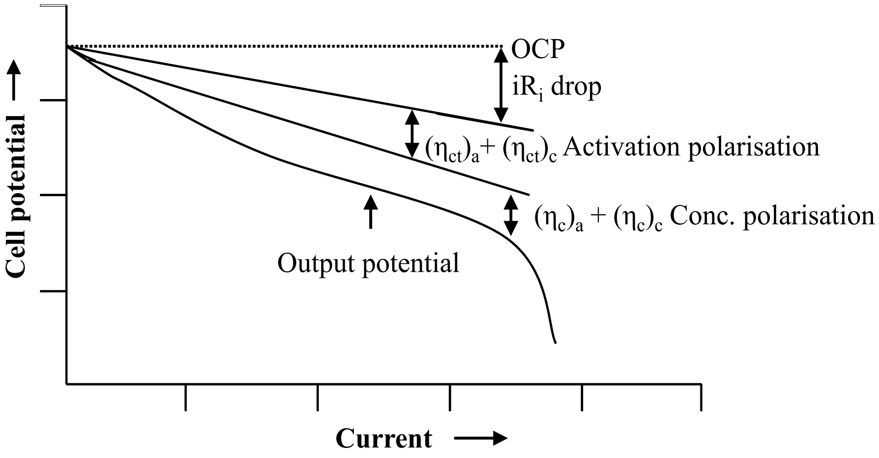


Figure 2. Cell potential as a function of current.

However, the available energy of a cell depends on the chemistry of the total system and principally the electrochemical reaction at both electrodes. There are some additional factors, which determine the kinetics of the charge-transfer reaction, the diffusion rate and degree of the energy loss. These factors include the electrode construction, cell engineering, electrolyte conductivity and nature of the separator, which will be discussed in detail in later sections. From a thermodynamic point of view, the reactions mainly occur at the two electrode interfaces, and during the discharge, the reaction at the cathode can be expressed by the following equation:

$$aA + ne \rightarrow cC \tag{2}$$

where a molecules of A are reduced to form c molecules of C, and n number of electrons e are consumed. The reaction at the anode can be expressed by

$$bB - ne \rightarrow dD \tag{3}$$

where b molecules of B are oxidised and converted to d molecules of D, and n electrons are released. The overall cell reaction is the sum of the above two half-cell reactions:

$$aA + bB \rightarrow cC + dD \tag{4}$$

The change in the standard free energy of this reaction is given by

$$\Delta G^o = -\,nFE^o \tag{5}$$

where ΔG^0 is the Gibbs free energy, F is the Faraday constant (96,485 Coulombs per mole) and E^0 is the standard potential which depends on the type of the active material integrated into the cell. This can be further represented for the non standard state condition of both electrodes by the Nernst equation:

$$E = E^o - \frac{RT}{nF} \ln \frac{a_C^c a_D^d}{a_A^a a_B^b} \tag{6}$$

where a_i is the activity coefficient of relevant species, R is the gas constant (8.31 $JK^{-1}mol^{-1}$) and T is the absolute temperature (K). The amount of electrical energy per mass or volume available for the external circuit is measured from the change in the standard free energy, the driving force of a battery. The total available energy in a cell is given by the following equation:

$$\Delta G = -\,xnFE \tag{7}$$

Where x is the molar quantity of the active materials involved during the discharge reaction.

Theoretically, the capacity of a cell is calculated from the quantity of the active materials and is expressed as the total quantity of electricity from the electrochemical reactions within the cell represented in terms of Coulombs or Ampere-hours. One molar mass or gram equivalent weight of the active materials involved in the electrochemical reaction produces 96,485 Coulombs or 26.8 Ah capacity. For example, the theoretical capacity of Li-ion battery can be calculated as follows:

$$\mathrm{Li_xC_6 + Li_{1-x}CoO_2 = LiCoO_2 + C_6} \tag{8}$$

x=0.5, for example and the capacity of the $LiCoO_2$ is 0.140 Ah/g

$$\text{Watt hour }(\text{Wh}) = \text{Cell Potential }(\text{V})\text{ x Ampere hour }(\text{Ah}) \tag{9}$$

As an example in the Li-ion battery with a standard potential of 4.1 V the theoretical Watt hour capacity per gram of the active materials (limited by the cathode capacity) is calculated as follows:

$$\text{Gravimetric Specific Energy }\left(\text{Whg}^{-1}\right) = 4.1\text{ V X }0.140\text{ Ahg}^{-1} = 0.574\text{ Whg}^{-1}\text{or }574\text{ Whkg}^{-1} \tag{10}$$

The practical energy density is decreased when the inactive materials, binders etc and the components of the electrolyte are included in the full cell calculations.

1.3. Rechargeable lithium batteries for energy storage

Li is attractive as the anode material for rechargeable batteries being the lightest metal (6.94 g mol^{-1}), with a standard reduction potential of -3.04 V (versus standard hydrogen electrode, SHE), resulting in the largest specific energy storage capability (3861 mAhg^{-1}). In the 1970s, Li metal was assembled in primary Li cells (non-rechargeable) for the first time, and because of their excellent capacity and discharge rate, they were used to power many electronic devices e.g. watches, calculators and implantable medical devices. [2] Around that time a number of inorganic compounds were also shown to have reversible-chemical reactivity with alkali metals opening up the possible use of intercalation compounds in lithium batteries. These materials were essential in the development of the high-energy Li-ion rechargeable batteries.

Early research used TiS_2, as the cathode material and Li metal as the anode material.[3,4] However, issues with the metallic Li anodes, in particular dendritic Li growth on cycling (which caused short circuits) raised safety concerns for the use of Li metal and resulted in the investigation of alternative approaches with new electrolyte and negative electrode materials. The alloying of Li with metals or intercalation of Li^+ into metals or semiconductors such as Si emerged as the potential solution. This solved the dendrite problem but cycle life was significantly influenced by large volume changes of the alloy material during the Li-ion insertion/de-insertion process. Initially, LiAl alloy with a composition of 1:1 was used as anode material with large specific energy capacity (780 mAhg^{-1}) but a volume expansion of 200% resulted in electrode crumbling, loss of electrical contact and rapid capacity fading. [5]

In the late 1970's and early 1980's researchers proposed the substitution of metallic Li with a second insertion material as the anode which led to Li-ion or 'rocking chair' batteries (Fig. 3). [6,7] Li was thus incorporated in the ionic rather than the metallic form in the anode electrode which solved the dendrite problem leading to safer Li batteries. The ionic character of the anode material (Li^+ intercalated material) increased the reversible anode potential and consequently a higher redox potential cathode material was needed to compensate for the anode. Layered or three-dimensional transition metal oxides, Li_xMO_2 (M= Co, Ni or Mn) which exhibit more ionic character than the transition metal disulfides became the cathode material of choice.[8,9] These commercially applicable cathode materials are still used to date in a large number of portable devices. [10,11]

The search for suitable Li ion anodes continued and after many years of extensive research carbonaceous materials with high reversibility were reported at a low Li^+ intercalation / deintercalation potential using liquid electrolyte at room temperature. [12,13] The discovery immediately led to the implementation of the rocking chair concept, and Sony Corporation, commercialised the C/$LiCoO_2$ Li-ion or rocking chair rechargeable battery with high energy density (≈ 180 Whkg^{-1}) and high discharge potential (3.7 V) in June 1991. [14] Li-ion rechargeable batteries are used in most of today's high-performance portable electronic devices.

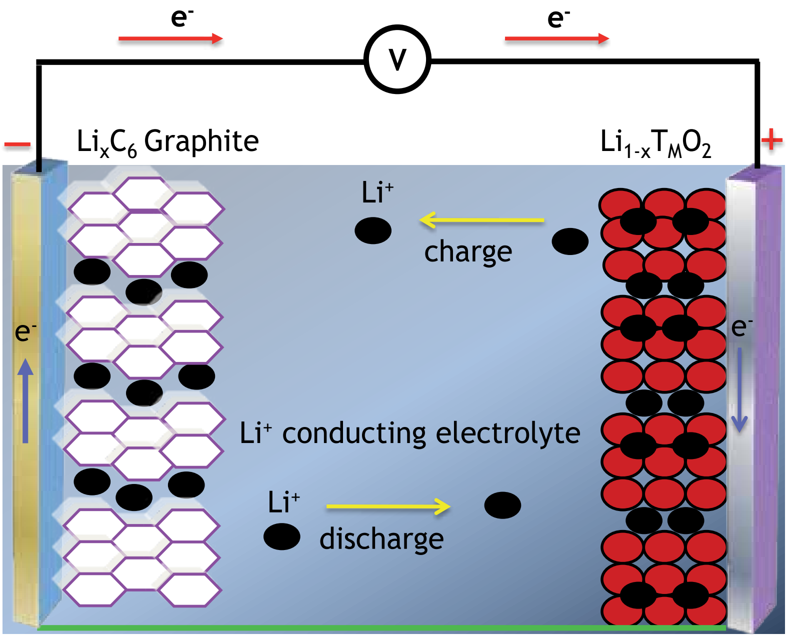


Figure 3. Representation of the lithium ion 'rocking-chair' rechargeable battery.

http://dx.doi.org/10.5772/57139

A further development investigated the use of polymer based Li-on rechargeable batteries, so-called Li-ion solid polymer electrolyte (Li-SPE) batteries, by substitution of liquid electrolyte with dry polymer electrolyte. [15] However, the poor conductivity of the SPE at ambient temperature (typically requiring temperatures above 80ºC) has prevented their widespread application in portable electronic devices. However, a new class of electrolyte was introduced by mixing a small amount of typical organic solvents with the polymers and lithium salt resulting in a polymer gel electrolyte. The plasticiser organic liquid mixed with the polymer results in an ionic conductivity approaching that of solvent alone. [16,17] Polymer Li-ion (PLiON) batteries operate at ambient temperature and offer a thin-film design flexibility. [18]

1.4. Electrode materials for Li ion batteries

Lithium ion batteries are in widescale deployment. Fig. 4 shows the gradual improvement in energy density for Li ion systems that has occurred over the 20 years since their introduction commercially in 1991. The improvements are attributable to improved engineering of the materials, their structuring, processing and additives utilised. An equivalent scale is included on the right hand axis to indicate an equivalent energy density for a 10 μm thin film version per cm^2 footprint if it were possible to scale all of the materials of standard lithium cells. The value is very similar to the actual thin film versions in use today which have a lithium metal anode rather than carbon.

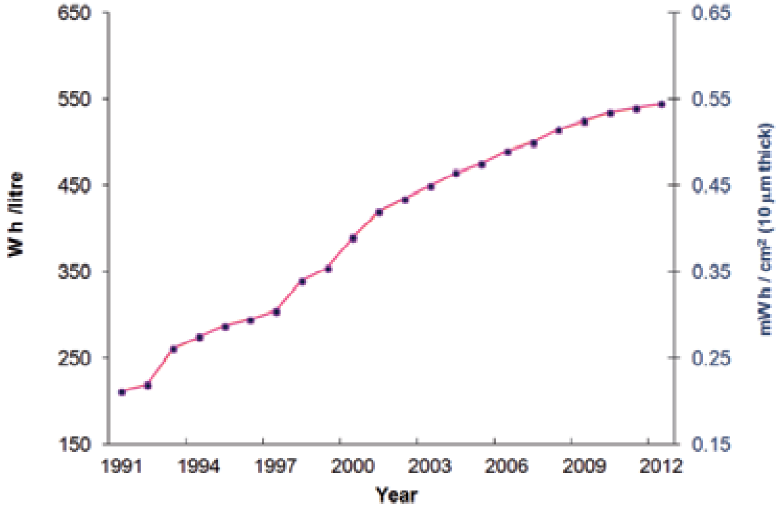


Figure 4. Li ion energy density improvements since 1991.

Carbon is used as the anode material in almost all commercially available Li-ion rechargeable batteries. Among the different types of carbon classified accordingly to their natural sources and structures, graphite (ABAB- layers) and hard carbon (polycondensation of oil pitch) are mostly used as anodes. The electrochemical intercalation has a Li capacity of 372 mAhg^{-1} from the corresponding lithiated formula LiC_6 at room temperature. [19-22] Due to the difference in structure hard carbon is able to incorporate Li^+ in the space between crystal particles as well as between the layers and as a result a higher capacity has been achieved in the hard carbon than the graphite within the same lithiated formula LiC_6. [23] The drawbacks of using carbonaceous materials as anodes include low specific energy capacity and electrolyte reactivity at the irregular surface.

The low intercalation potential, only 0.3 V away from that of Li, enables 3.5 to 4 V lithium batteries though the overall energy is approximately 10 times less than if a Li metal anode was employed. The proximity to the potential for lithium metal plating means that care must be taken in the operation of the battery to prevent overcharge and Li metal plating during which dendrites can form that penetrate the electrolyte separator and short the battery. Modifications are continuously being introduced and a few metals, e.g., Ag, Zn or Sn, coated graphite fibre have been reported as high charge/discharge capacity anode materials. [24] It is believed that the improvements afforded by the coating result from modified passivation films formed on the metal surface, which reduce the extent of reaction between the bare graphite and the electrolyte. At the same time research has been focused in the development of alternative anode material with both higher capacities and slightly higher intercalation potential compared to carbon/Li^+ and Li/Li^+ in order to avoid Li plating on the high rate charging.

1.5. Anode materials

Materials with advanced performance such as Li metal alloys have been considered as potential alternatives for carbonaceous anode materials. Although attractive in terms of higher gravimetric capacity, the cyclability has generally been very poor in deep discharge, due to large volume changes (up to 200%) during Li^+ insertion and de-insertion, which causes mechanical disintegration and hence loss of electrical contact between the active material and the current collector. [25] To overcome this issue, attempts have been made to introduce an inactive phase (buffer matrix) into the Li alloy to suppress or compensate for the volume expansion in some extent, while still protecting the electrical pathway. [26] Such alloy systems are achieved by mixing two or more metals so that electrochemically active metal phases are embedded in electrochemically inactive phases to react with Li forming alloys. Recently, Sn based alloy compounds Sn-M′, where M′ is an electrochemically inactive confining buffer have been studied extensively because of their higher specific capacity, such as Sn-Fe[27], Sn-Co [28], Sn-Zn-Cu [29] and Sn-Ni. [30] However, metal alloys still suffer from mechanical strains during Li insertion and de-insertion that leads to cracking and crumbling of the electrode. These alloys show much better cyclability over simple Li alloys at the cost of reversible capacity; consequently, further improvement is needed for the practical application in Li-ion batteries.

An approach to reduce the problems of volume changes by selecting intermetallic alloys, such as Cu_6Sn_5, InSb and Cu_2Sb has also been considered because of strong inter-structural relationship between the parent compound and the respective lithiated products, Li_2CuSn, Li_3Sb. [31] Research has been focused on those alloys because they react with Li topotactically, which has a structure of cubic symmetry in the initial transitional phase that provides a stable host framework for both the incoming and extruded metal atoms. They expand isotropically and to a lower extent than most alternatives. In spite of this they still suffer from relatively poor cyclability due to electrical and mechanical disintegration and require further optimisation.

Another route to alleviate the poor cyclability of the metal alloys is to use nanocomposites of active and inactive materials. [32] Several reports have been published on nanocomposite anode materials, for example SnO based glasses [33], Sn-Fe-C [34], Sn-Mn-C [35], Si-C [36], and Sn-Co [37], which show improved cycling behaviour. It is argued that the nanosized metallic clusters suppress the associated strains, and thus enhance the reversibility of the alloying reaction. However, nanocomposites still exhibit large capacity losses in the first cycle and capacity fading in subsequent cycles. This is related to a combination of irreversible processes, such as aggregation of nanoparticles, irreversible trapping of Li^+ by host clusters and secondary reactions involving electrolyte decomposition and the formation of unstable passivation layers. Further strategies have been investigated to minimise such secondary electrolyte reactions by uniformly coating the alloy composite surface with a less reactive protecting layer. [38]

More recently hollow nanospheres of various materials including metals, oxides and semiconductors have been considered as potential anodes. [39-41] Preparation methods for nanospheres with hollow interiors typically involve the removal of sacrificial templates, including silica [42] and polymer latex spheres [43], or reducing metal nanoparticles. [44] The hollow Li active nanospheres serve as barriers to particle pulverisation and provide a large surface area to buffer the volume change in Li insertion/de-insertion, consequently improving the capacity retention. Hollow nanospheres of Sb with substantially superior capacity retention than its counterpart nanoparticles at higher rates have been reported. A schematic illustration of the Sb hollow nanospheres typical of nanosphere preparation routes investigated is shown in Fig. 5 below. [40] Among metal oxides, SnO_2 has been widely studied with a theoretical specific Li storage capacity (ca. 790 $mAhg^{-1}$). The main obstacle in commercial application of the Sn based anode materials is the pulverisation that occurs on cycling. [45,46] Nanoparticles combined with elastic hollow carbon spheres such as Sn nanoparticles encapsulated within conductive carbon nanospheres have also been investigated. [47] The free space inside the shell left after materials loading (void volume) and the elasticity of thin carbon spheres accommodate the strains associated with Li^+ insertion/de-insertion, as the volume can expand or contract repeatedly.

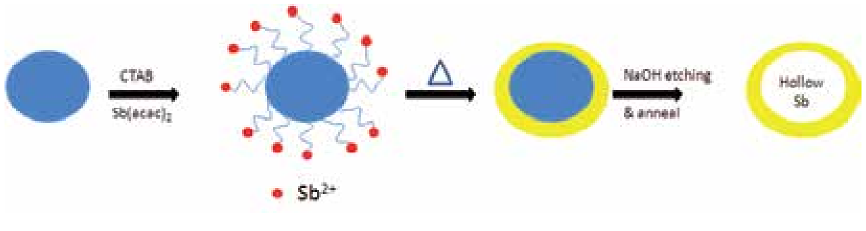


Figure 5. (a) Schematic illustration of the preparation of hollow Sb nanospheres from the CTAB (cetyltrimethylammoniumbromide) functionalised SiO_2 templates.[40]

2. Cathode materials

The materials for the cathode are typically selected according to the anode utilised. In rechargeable batteries with metallic Li as anode, the cathode does not need to be lithiated before cell assembly. On the other hand, in Li-ion batteries where the anode is made of Li active metals, alloys or carbon, the cathode should act as a source of Li^+ and thus involve the use of air stable Li intercalated compounds. $LiCoO_2$ has been used as the cathode material since the first commercial Li-ion batteries were introduced, due to its structural stability (in limited cycling), ease of preparation and higher discharge potential than the counterpart dichalcogenides (TiS_2). [14] However, the specific capacity of the $LiCoO_2$ is limited to only 140 $mAhg^{-1}$ at room temperature for 0.5 Li/Co, although the desired theoretical value is 273 $mAhg^{-1}$. [48] The reduced specific capacity is associated with the restricted range of reversible cycling as almost delithiated Li_xCoO_2, when $x < 0.3$, decomposes the electrolyte and causes electrode corrosion. The $LiCoO_2$ delivers the best power density if discharged between 4.2 V to 3.0 V vs. Li/Li^+. $LiCoO_2$ continues to be used in commercial batteries that power cell phones, laptops, etc., despite higher cost and both environmental and safety issues. But these issues limit their application where low cost and higher energy capacity is required, such as hybrid electric vehicles (HEV) or electric vehicles (EV).

Lithium nickel oxide, $LiNiO_2$, which is isostructural with $LiCoO_2$ has been considered because of slightly higher specific capacity, lower cost and much lower redox potential that diminishes electrolyte oxidation although safety concerns must also be addressed for this material most likely by the combination with additional metal oxides. $LiNi_{1-x}Co_xO_2$ systems have been found to be more stable than the Li_xNiO_2, but further stabilisation of that layered structure is essential for safety and capacity issues. Several routes have been investigated and the addition of a redox-inactive di-, tri- or tetravalent cation (Al, Ga, Mg or Ti) substitute for Ni or Co appears to be the most promising approach. [49-51] The substituted phases, such as $LiNi_{1-x-y}Co_xAl_yO_2$ are reported to be safer cathode materials for applications that require large energy capacity. [52] The substituted inactive-element prevents Li from deintercalating entirely and thus maintains the O_2 partial pressure at a level that prevents possible structural collapse of the delithiated phase. SAFT incorporated these cathode materials for the first time in practical Li-ion batteries realising an energy density of 120-130 $Whkg^{-1}$ even in deep discharge conditions. [53]

Layered $LiMnO_2$ is another popular choice for the next generation Li-ion batteries not only as an abundant material source but also because it is environmentally benign. [54] The structure differs from that of $LiCoO_2$ and $LiNiO_2$ consisting of alternating zigzag layers formed from corner sharing between LiO_2 and MO_6 octahedra. $LiMnO_2$ shows a specific charge capacity of 190 $mAhg^{-1}$ in the potential range of 2.0 V to 4.25 V with 99.9 % capacity retention. Unfortunately, it follows the same fate of capacity fading upon delithiation owing to the presence of extra negative charge in the MnO_2 layers. The structural instability of the layered Li_xMnO_2 phase has led to the investigation of the spinel $Li_xMn_2O_4$ structure with 3D channels which allow faster Li^+ diffusion upon cycling. [55]. Its implementation has been delayed mostly due to poor cyclability and low stability owing to dissolution of Mn^{+3} into the electrolyte at elevated temperature, which occurs at the particle surface in the presence of trace acids, especially HF. The dissolution of Mn^{+3} results in the formation of a defective spinel $LiMn_2O_4$ that reduces Li $^+$ insertion capacity. Partial substitution of the Mn by more electron rich elements, e.g., Co [56, 57] or Ni ($LiMn_{1-x}Ni_xO_2$, $0 < x \leq 0.5$) [58,59] has also been investigated. Although the initial specific capacity of these materials is as high as 200 mAhg^{-1} they show poor cyclability upon overcharging. The role of the Mn is to stabilise the layered structure of NiO_2, and the substituted Ni is redox and electrochemically active between the Ni^{+2} and Ni^{+4} states. The substituted Co within the structure controls the 3D channel in the spinel phase. More recently, the solid solutions of these lithiated transition metal oxides $LiMn_{1-x-y}Ni_xCo_yO_2$, have been shown to overcome structural instability although electrochemical performance of these materials largely depends on the synthesis techniques, conditions and stoichiometry. [60-66]

Phospho-olivine $LiFePO_4$ phase has also been introduced commercially because it is plentiful, environmentally benign, low-cost, non toxic and has a competitive theoretical capacity of 170 mAhg^{-1}, higher than that obtained for layered $LiCoO_2$ and $LiNiO_2$. [67-69] The main disadvantage associated with $LiFePO_4$ is poor electronic conductivity and consequently, significant efforts have been made to overcome this through chemical and physical processes. Carbon coatings on small $LiFePO_4$ particles [70-72] or $LiFePO_4$/PPy [73] (polypyrrole) composites have been utilised to enhance the capacity and rate capability of $LiFePO_4$. Carbon coated porous $LiFePO_4$ works well at low discharge rates but not at high discharge rates, due to the inadequate electronic conductivity of the material. Recently, the issue of poor electronic conductivity of porous $LiFePO_4$/C composite has been addressed by introducing oxidic (RuO_2) nanoscale interconnects. [74] The RuO_2 oxide coating repairs the gap between carbon conducting networks on porous $LiFePO_4$ and thus improve the kinetics and rate capability of the $LiFePO_4$/C composite.

Vanadium pentoxide, V_2O_5, and its derivatives (Cr, Nb and Mo) were of the earliest studied cathode materials for their high specific capacity, low cost and ease of extraction from minerals. [75-77] The vanadium oxide family as well as dichalcogenides of transition metals MX_2 (Ti, Nb, Ta, Mo and W) have a CdI_2-type structure, with disordered close-packed chalcogen layers between which the transition metals reside in either prismatic or octahedral coordination of six chalcogens. The theoretical capacity of V_2O_5 is the highest among other family members (442 mAhg^{-1}). However its rapid potential change with degree of Li^+ insertion that passes through different phases and low reaction rate makes it less manageable for practical applications. [78] At the same time a number of vanadium oxides, such as partially reduced V_6O_{13} [79] and LiV_3O_8 [80] have also been studied with the conclusion that the electrochemical performance depends on the synthesis technique. [81]

To summarise, lithium ion cathode materials fall into two groups in general, the first with more compact lattice structure, such as $LiCoO_2$, $LiNO_2$, $LiMnO_4$, substituted lithiated transition metal oxides ($LiCo_{1-x}Ni_xO_2$, $LiNi_{1-x}Co_xO_2$ or $LiMn_{1-x}M_xO_2$, M= Ni or Co) or their solid solutions ($LiMn_{1-x-y}Ni_xCo_yO_2$), and the second with more open structure like V_2O_5, MnO_2 and olivine phase ($LiFePO_4$ or $Li_3V_2(PO_4)_3$). But most of the cathode materials exhibit low electronic conductivity ranging from 10^{-3} S cm^{-1} for $LiCoO_2$ [82] down to 10^{-9} S cm^{-1} for $LiFePO_{4.}$ [83] which is fourteen orders of magnitude lower than Cu (5×10^5 S cm^{-1}) and thus surface or substrate modification is required to alleviate this problem. To improve the Li^+ insertion and extraction kinetics, reducing the size of such classical cathode materials is also necessary to achieve the short diffusion length and large contact area for higher current drain though it does also expose a greater area for unwanted electrolyte interaction. A nanosize coating that covers the whole surface of nanostructured active materials without contamination and aggregation is a great challenge. 1D nanowire and nanotube arrays or 3D macrostructures would be advantageous over nanoparticulate composites, and are discussed below. Li battery electrode materials and their electrochemical potential and capacity are summarised in Fig. 6.

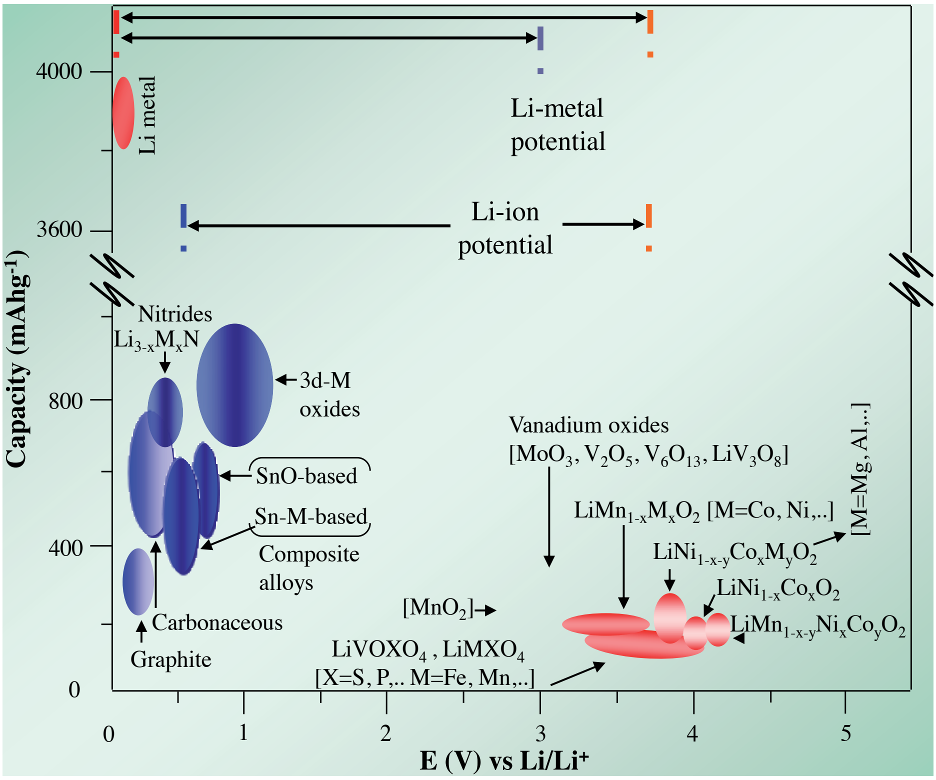


Figure 6. Schematic illustration of materials for negative and positive electrode in terms of capacity and potential currently used or projected for the next generation of rechargeable Li-ion or Li-metal batteries.

3. Electrolytes

In Li-ion battery systems organic solvents with a large stability window, good ionic conductivity, low melting and high boiling point and a low vapour pressure are required (see Table 1). Many studies have been carried out to choose suitable electrolytes for rechargeable lithium batteries based on thermal stability of the electrolytes [84-86] and electrochemical kinetics of the lithium ion in different mixed solvent electrolytes systems of various compositions. [87-88] The currently used cyclic carbonic acid esters like ethylene carbonate (EC) have high dielectric constant but high viscosity values due to the interaction between molecules which hinders mass transport of solutes. High dielectric constant solvents have high coulombic force between positive and negative sites in a molecule which results in a degree of ionic dissociation for the complex lithium salts. However, these cyclic carbonates show improved performance when mixed with a chain-like ester such as diethyl or dimethyl carbonate (DEC, DMC) whose viscosity and dielectric constant values are quite low. These solvents are commonly known as thinning solvents and it is for this reason that the solid phase of EC at room temperature is usually blended with these solvents.

Solvent	Dipole Moment	Dielectric Constant	Melting Point	Boiling Point	Density	Molar Volume
	(μ/D)	k or (ε)	(^{0}C)	(^{0}C)	(g/cm^3)	(dm^{-3}/ mol)
Ethyl carbonate (EC)	4.9	89.78	37	248	1.32	62
Propylene carbonate (PC)	4.94	66.14	-49	242	1.2	84
Diethyl carbonate (DEC)	0.94	2.82	-43	126	0.97	122
Dimethyl carbonate (DMC)	0.88	3.12	3	90	1.06	84

Table 1. Physical properties of organic solvents at 25^{0}C

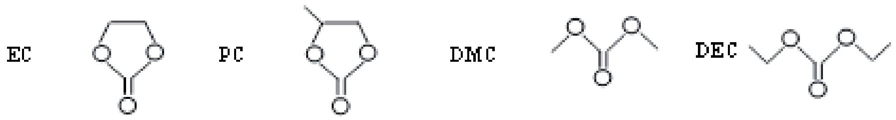


Figure 7. Structural formula of main organic solvents.

Although, LiFP$_6$ electrolyte salt is stable in the above electrolytes it produces a strong Lewis acid, PF$_5$, in the presence of trace amount of water by the following mechanism. [89] PF$_5$ attacks the lone pair of electrons on the oxygen in water molecules and decomposes.

$$LiPF_6 \rightarrow LiF + LiP_5 \downarrow \tag{11}$$

$$LiP_5 + H_2O \rightarrow PF_3O + 2HF \tag{12}$$

Thermal decomposition of 1M LiPF$_6$/EC:DEC at 85^{0}C is similar to the reaction between EC/DMC and PF$_5$ gas. [90] Strong Lewis acids cleave the EC ring and produce transesterification products. Ethyl groups are electron donating, which is why PC is more reactive with

PF_5 and decomposes more easily than DEC. Since PC based mixed electrolyte is reported to decompose in contact with graphite anodes, an appropriate aprotic solvent mixture of EC/DEC has been recommended. [91-92]

It has been reported that the diffusion coefficient of EC based mixed electrolytes increases with increasing solute concentration though the solution viscosity increased with higher solute concentration. [15] The Li ion diffusion coefficient of the organic electrolyte systems cannot be directly extrapolated to their viscosity in terms of solute concentration, as other factors contribute such as ion solvation ability, ion conductivity, mass transport, solvent-solvent interaction, reduction and absorption of solvent on the electrode surface. The organic solvents PC and EC (Table 1) have nearly the same value of dipole moment and permittivity which is comparable with those values of water (1.855 D, 81 ε). The high solvation effect in organic electrolytes compensates for the increase of viscosity with increasing salt concentration.

The diffusion coefficient values for Li^+ increases in the order of $PC/LiClO_4$ < EC: DEC (1:1)/ $LiPF_6$ < EC: DMC (1:1) /$LiPF_6$ with an increase from 0.1M to 1M in solute concentration.[93] The maximum diffusion coefficient values for lithium ion are $1.2x10^{-5}cm^2/s$ and $1.39x10^{-5}cm^2/s$ for 1M $LiPF_6$ in EC: DEC (1:1) and EC: DMC (1:1), respectively. In EC based mixed electrolytes lithium ion is coordinated by the EC rather than the acyclic carbonate (DEC, DMC) owing to high polarity and slightly higher donor number of EC (DN, 16.4 for EC, 15.1 for DEC and DMC). Strongly solvated lithium ion complexes with high permittivity EC and reduces the ion association effect in low viscosity DEC or DMC. EC based mixed electrolytes of 1M $LiPF_6$/EC/ DEC or 1M $LiPF_6$/EC/DMC have become the liquid electrolytes of choice for use in lithium rechargeable batteries.

To minimise safety issues in Li-ion rechargeable batteries all solid electrolyte are very attractive. Armand et al. first investigated the ionic conductivity of Li^+ in solid state polyethylene oxide (PEO) polymer and applied it to Li-ion rechargeable batteries. [15]. These electrolytes are intrinsic solid polymer electrolytes (SPE) and the crystallinity favoured at lower temperature decreases the ionic conductivity. To achieve higher conductivity electrolytes at room temperature polymers were mixed with small amounts of the typical aprotic solvents known as plasticiser, such as, EC, PC, DEC and DMC. [93]. Polymer-gel electrolytes consist of polymer networks swollen with liquid(s) and possess both the cohesive properties of solids and ionic transport properties like liquids. Polymer gel electrolytes are alternatively called "polymer hybrid" or "gelionics".

Polymer gels containing alkali metal salt trapped within the matrix of the polymer host were first demonstrated by Feuillade et al. in 1975, with an ionic conductivity close to that of the liquid electrolytes. [94] Afterwards, polymer gel electrolytes with a variety of polymer hosts, such as polyethylene oxide (PEO) [95] polyvinylidene fluoride (PVdF) [96,97] polyacrylonitrile (PAN) [98,99] polymethyl methaacrylate (PMMA) [100,101], polyvinylidene fluoride-hexafluoropropylene (PVdF-co-HFP) {102,103] have been reported with the ionic conductivity values in the range of 10^{-4} to 10^{-3} S cm^{-1} at ambient temperature. The plasticisers increase the amorphous character of the host polymers with a single glass transition temperature (Tg) below -40°C. Thus, the ionic conductivity of gel electrolytes increases through enhanced diffusive transport in the liquid phase. PAN and PVdF-based polymer gels are the most widely studied electrolyte systems. PAN-based gel electrolytes with dispersed lithium salts, for instance, $LiClO_4$, $LiAsF_6$ and LiTFSI, have been shown to have very high ionic conductivity and Li^+ transfer number. [104] However, these electrolyte systems could be applied in batteries only with intercalation electrodes, because of the high reactivity with Li metal.

PVdF has enhanced capability to dissociate Li salts due to the strong electron withdrawing functional group (-C-F) in the host structure as well as a high dielectric constant (ε=8.4). Cast-polymer-gel electrolyte films based on PVdF solid polymers mixed with plasticisers EC/PC and Li salts, for example, $LiCF_3SO_3$, $LiPF_6$ or LiTFSI have been prepared by Jiang et al. [105] The mechanical stability of these PVdF-based gel electrolytes varied with the amount of the solid polymer and the concentration of the dispersed Li salts determined the total ionic conductivity. However, PVdF-based gel electrolytes are also unstable towards Li metal and Li salts due to the fluorinated polymer host, which produces poor interfacial characteristics. The co-polymerisation of PVdF with hexafluoropropylene (HFP) has been reported to enhance the electrolyte properties significantly, giving higher solubility rates in organic solvents and lower crystallinity with a reduced glass transition temperature (Tg) than pure PVdF in the gel. {106]. Although, the polymer-gel electrolytes have high ambient ionic conductivity the mechanical strength is decreased and a higher interfacial impedance is observed due to the passivation layer at the Li/Li^+ interface. [107-110]

Recently, the problems associated with the polymer-gel electrolytes have been minimised by introducing nanocomposite gel polymer electrolytes instead of conventional/gel/plasticised polymer electrolytes, but at the cost of a slight decrease in the conductivity value. The dispersion of nano-ceramic filler particles (~10% w/w), such as Al_2O_3, SiO_2, and TiO_2, into polymer gel electrolytes increases the mechanical stability, where nanoparticles act as physical and chemical barriers to the solvents evaporation from the system. [111] The technologies based on either solid or hybrid polymer electrolytes offer great advantages to meet the shape and design flexibility requirements for miniaturisation of electronics and other portable devices. Additionally, solid Li-polymer technology enables excellent packaging efficiency without the possibility of solvent leakage.

4. Energy storage materials and architectures at the nanoscale

Nanotemplated materials have significant potential for applications in energy conversion and storage devices due to their unique physical properties. Nanostructured materials provide additional electrode surface area with short path lengths for electronic and ionic transport and thus the possibility of higher reaction rates. [112-114] Control of the active materials at the nanoscale is required for battery materials where solid-state ionic diffusion is a limiting factor in the electrode reactions. Template mediated fabrication is a potential method for ordered high density array fabrication of nanowire/nanotube on metallic current collector substrates. Early reports of nanotemplated materials focused on the use of polymer templates. [115-117] Subsequently, higher pore density anodic aluminium oxide (AAO) substrates (10^9 to $10^{11}/cm^2$ by comparison with $10^8/cm^2$ for track etched polycarbonate) have been developed. [118] It is also possible to form pores with smaller diameter (to low 10's of nm level) vertically aligned for more ordered channels and active materials. AAO processing on Si has been investigated [119-121] and is under development as a means to optimise the materials nanotemplating and integration of passive devices with silicon technology. [122,123] Other fabrication methods include vapour-liquid-solid (VLS) growth. [124] VLS growth combined with chemical vapour deposition (VLS-CVD) [125] and template mediated micelle deposition. [126] Optimising active materials and architectures not only requires assessment of the electrochemical properties for the given application but also the electrical and mechanical characteristics during reaction.

1D nanowires or nanotubes have anisotropic morphologies and self-supported arrays grown directly on a current collector represent an attractive architecture for Li-ion batteries. Such arrays can ensure that each nanowire or nanotube participates in the electrochemical reaction with 1D electron transport pathways. Arrays of nanowires and nanotubes eliminate binders or additives that decrease power density by incorporating extra inactive materials in the total weight. Moreover, they can accomodate strain during charge and discharge. [127,128] Nanotubes function as electrolyte-filled channels for faster ionic and mass transport to the remote electrode surface. A variety of synthesis methods for the preparation of 1D nanowire or nanotubes for use as Li battery electrode materials have been investigated including sol-gel processing combined with template synthesis or hydrothermal treatment. The following is a brief review of recent progress in advanced anode and cathode materials for use in Li-ion batteries.

4.1. Anode materials at the nanoscale

A number of metals and semiconductors, for instance, Al, Sn, Si, accommodate a large number of Li atoms per formula unit in electrochemical reactions, and provide higher specific capacity than that offered by the conventional graphite. Unfortunately, large volume changes and phase transitions in the host metal accompany the reactions. Attempts to limit the mechanical strains generated during Li ion insertion and removal were discussed in the earlier section including the use of active/inactive nanocomposites, hollow nanospheres and nanoparticles encapsulated with elastic shells. Although, nanocomposites or nanospheres considerably suppress the associated strains, and thus improve the reversible capacity, unsolved issues include low energy density of the electrode due to the poor packing density and the large proportion of inactive binders, additives, etc., in the electrodes. [129]

1D nanowires or nanotubes of metals, semiconductors or oxides provide a route to enhanced battery materials as they suppress strains, minimise pulverisation and provide good electronic contact and conduction pathways when grown directly on the current collector. Cu_6Sn_5 nanowires [130] with a measured height of 5 μm and diameter of 250 nm or an aspect ratio of 20 have been studied by cyclic voltammetry, figures 8 and 9. The first charge and discharge capacity of the nanowires was 325 mAh/g. The charge–discharge capacity is stabilised at 175–200 mAh/g from the third cycle with no significant deterioration. The obtained results imply that a topotactic reaction mechanism for Li^+ insertion/extraction is established with suppressed irreversible capacity loss due to the non-transferable lithiated transition phase and oxide impurities. [131, 132]

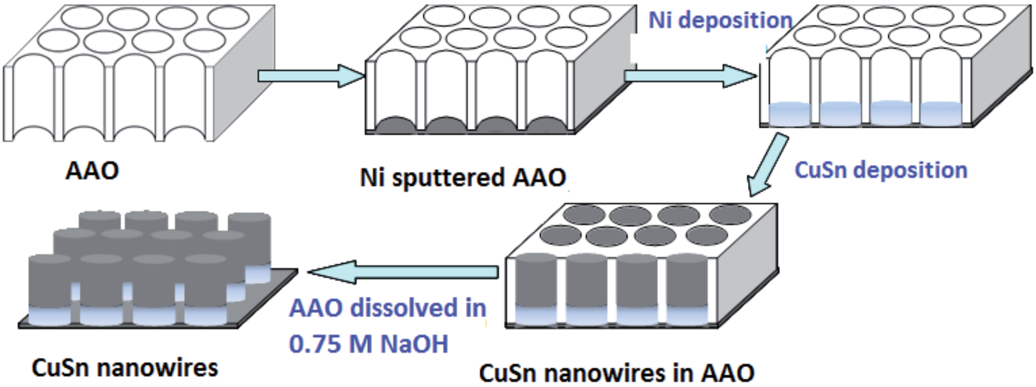


Figure 8. Schematic of templated CuSn nanowire fabrication process using Ni backed AAO.

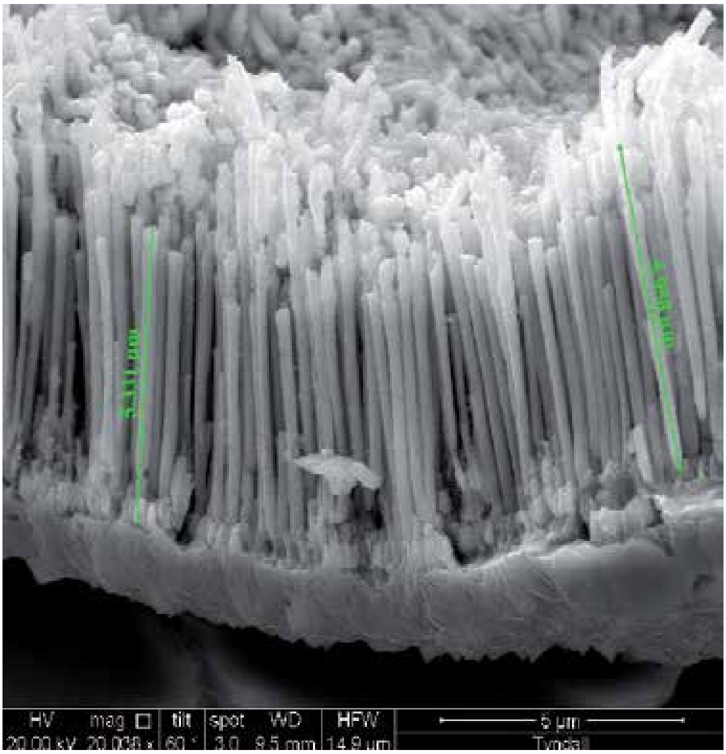


Figure 9. CuSn nanowires grown on Ni supports in an AAO template.

Alternative high capacity anode materials that have received attention as potential replacements for carbon and that may be processed on Si for micro energy storage include, Sn (990 mAh/g) Ge (1,600 mAh/g) and Si (4,200 mAh/g). In each case the materials can intercalate up to 4.4 moles of Li per mole and as a result suffer significant expansion and contraction on lithium cycling. Whitehead et al [133] reported that mesoporous Sn may be deposited from lyotropic liquid crystal electrolytes and that the mesoporous Sn exhibited enhanced capacity and a slower loss of capacity than non templated Sn. The capacity did, however, drop below 400 mAh/g within 10 cycles. Vertical arrays of one-dimensional Sn nanowires on silicon dioxide (SiO_2)/silicon (Si) substrates have been developed as anode materials for lithium rechargeable microbatteries. For these 1 D nanowires a discharge capacity of 400 mA h/g could be maintained after 15 cycles at the high discharge/charge rate of 4200 mA/g. [121] Ge semiconductor has a theoretical capacity of 1600 $mAhg^{-1}$, significantly higher than C but less than half that offered by Si. It has the advantage of a 400 times higher diffusivity for Li than Si. [134] Ge nanowires were fabricated by a template-free VLS method combined with CVD on the same stainless steel substrate using Au as catalyst and found to have initial discharge capacity of 1141 $mAhg^{-1}$ at the 0.05 C rate. [125] The capacity dropped to approximately 600 mAh/g at higher rates but with these values retained to at least 20 cycles indicating that nanostructured anodes can retain capacity despite large volume changes on cycling.

1D SnO_2 materials are one of the most extensively researched nanostructured anode materials for high-energy density Li-ion batteries. [135-137] Recently, highly ordered amorphous-CNT coated single crystal SnO_2 nanowire arrays have been fabricated by drying and annealing the SnO_2 sol-gel filled AAO template using citric acid as chelating agent. [127] These amorphous-CNT coated single crystal SnO_2 nanowires have shown reversible capacity of 418 $mAhg^{-1}$ in the first cycle and capacity retention of 353 $mAhg^{-1}$ after 30 cycles. Nanotubes of the polycrystalline SnO_2 have been synthesised through an infiltration technique using SnO_2 nanoparticles as starting building units and AAO as template. [138] The tubular templated SnO_2 nanoparticles have demonstrated a significant improvement of the specific capacity and the cyclability over their non-templated nanoparticle counterparts, and a reversible capacity of 525 $mAhg^{-1}$ was retained after 80 cycles. A similar approach of using SnO_2 nanotubes as high performance anode materials has been reported where nanotubes were synthesised first, and then a uniform CNT over-layer was grown on the external surface of the SnO_2 nanotubes through an AAO template assisted space-confined catalytic decomposition process. [139] The SnO_2-core-C-shell nanotubes obtained have been shown to deliver a higher reversible capacity of about 600 $mAhg^{-1}$, and excellent cyclability with capacity retention of 92.5% after 200 cycles. The enhanced electrochemical performances were assigned to the unique features of SnO_2 nanotubes, including flexible thin walls that accommodate strains uniformly caused by electrode volume change, and open ends allowing more efficient Li^+ insertion and transportation.

The electrode reactions using classical intercalation materials involve the insertion (or de-insertion) of Li^+ into (or from) an open host structure with a concomitant addition or removal of electrons. Taberna et al., extended the idea of the so called "conversion reactions" to 1D Fe_3O_4 nanowire structures, the reaction mechanism is different from the classical Li^+ insertion/

de-insertion. [140] High capacity and rate capability, for example, 80% capacity retention at 8C rates, have been reported using these nanowire arrays. The specific capacity of Fe_3O_4 is considered about 928 $mAhg^{-1}$ by assuming the reduction of the host metal ions F^{+3} or Fe^{+2} to Fe^0 during Li^+ interaction. Mesoporous Co_3O_4 nanowires of the same "conversion reactions" family have been synthesised by an ammonia-evaporation-induced method on various substrates such as Si wafer, glass slide, Cu or Ti foil, and polystyrene. [141] This is a mild template-free method, which allows the growth of large area nanowire arrays and many choices for substrate. The mesoporous Co_3O_4 nanowires have shown high specific capacity, good cyclability and high rate capability. They are reported to maintain a stable capacity of 700 $mAhg^{-1}$ at 1C rates after 20 cycles but at higher rates the capacity decreased to 85%, 69% and 50% when discharged at 8C, 20C and 50C, respectively. The main issue with such conversion reaction anodes is the large hysteresis between charge and discharge and the resulting relatively high cut-off potential, which can be as much as 3V vs. Li/Li^+, required to achieve full capacity. In this laboratory we fabricated Cu nanotubes [142] and converted the outer surface to Cu_2O though the use of an oxygen plasma ash for use in lithium batteries, figure 10. The oxide shell thickness is readily controlled exhibiting a linear relationship with time. [143] Over-oxidising the Cu nanotubes led to a poor cyclability (capacity loss within 10 cycles) for the anodes based on the loss of mechanical support for the active materials and the high conductivity Cu core. On the other hand when a core-core was maintained the electrodes cycled without loss of capacity for more than 90 cycles. This core-shell processing is therefore worth exploring to develop alternative anodes that do not suffer from the large hysteresis currently experienced by conversion anodes.

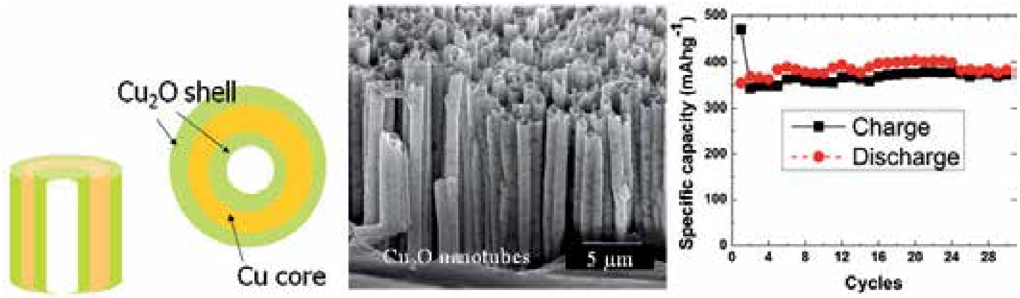


Figure 10. Cu_2O nanotubes for core/shell battery anode materials.

Si nanowires, are one of most attractive anode materials for Li-ion batteries and have also been synthesised by the template free VLS method on stainless steel substrate using Au as catalyst, and they are found to demonstrate higher capacity (4,277 mAh g^{-1} at very low C/20 rates) than other structured forms of Si. [144] In a subsequent paper [145] they showed that control the potential limit for Li intercalation had a significant impact on cycle life through suppression of the crystallization of a-Si of the Si nanowires. The capacity retention was observed to decrease steadily after 10 to 20 cycles even with the control of the lower limit for cycling. To overcome the issue of capacity fade that group succeeded in fabricating core-shell, crystalline-amorphous Si anodes [146] which achieved significantly improved cycle retention at the cost of lower energy density of apprximately 1000 mAh/g for Si. Recently, mesoporous Si-C core-shell nanowires have been prepared by Cho et al. using highly ordered mesoporous SBA-15 silica templates consisting of hexagonal arrays of 2D parallel cylindrical pores. [147]. The mesoporous Si-C nanowires have demonstrated initial charge specific capacity 3163 $mAhg^{-1}$ and a capacity retention 87% after 80 cycles at low 0.2 C rates. Such core-shell anode materials are of significant interest for future battery anodes and optimised materials and structures are required that will facilitate the high energy density coupled with high rate capability and with a footprint dimension no larger than the electronics they power.

4.2. Cathode materials at the nanoscale

A number of reports of 1D nanowire and nanotube transition metal oxides as Li ion cathodes have been published. Martin et al. investigated polycrystalline V_2O_5 nanowires (70 nm) which delivered significantly higher specific capacity than micrometer-sized diameter electrodes at low temperature. [148, 149] 1D battery electrodes meet the challenge for working at low temperature by confining the dimensions of the electrodes for enhanced Li ion diffusion. Cao et al. reported a template based electrodeposition method for single crystalline V_2O_5 nanowires by applying an electric field around the solution or sol. [150,151] Such nanowire arrays have been reported to possess higher capacity and higher rate capability than the polycrystalline thin film counterparts. Mixed valency vanadium oxide nanotubes (VO_x-NTs) have been prepared using a sol-gel reaction, followed by a hydrothermal step, from vanadium alkoxide precursor and primary monoamines that function as molecular structural directing template. [152] Wang et al. prepared nanotube arrays of amorphous V_2O_5 using the template based electrodeposition method at lower voltages and shorter deposition times than the conditions for preparing nanowires. [153] The initial Li capacity of the V_2O_5 nanotubes is reported to have specific capacity of 300 $mAhg^{-1}$, about two times higher than that of the V_2O_5 film (140 $mAhg^{-1}$), resulting from the increased surface area and smaller Li^+ diffusion paths. The reported capacity for the nanotubes after 10 cycles, of 160 $mAhg^{-1}$, was 30% higher than thin film electrodes.

Despite the excellent electrochemical performance of the traditional layered type V_2O_5 cathode material that accommodates intercalated Li^+ between the interlayers, there are also some other oxides showing high Li storage capacity by electrochemically reacting with Li^+ ions. West et al. prepared freestanding nanowire arrays of amorphous MnO_2 by electrodeposition into anodised alumina membranes and these are capable of multiple charge-discharges with specific capacity of approximately 300 $mAhg^{-1}$ in a half cell reaction. [154] Despite the relatively large surface area the electrodes could only be discharged at 100 $\mu A/cm^2$ before significant polarisation losses were observed.

A comparison of micron scale and nanoscale MnO_2 (figure 11 (a) material as a lithium ion cathode is shown in figure 11(b). Data is included for micron scale commercial MnO_2 powder. Also shown is the data for the 18 μm in length 250 nm diameter MnO_2 nanowires on Cu supports. In the first cycle for the nanowires an improvement is observed for the peak position and rate capability of the lithium insertion and extraction with respect to the micron scale material. The influence of the copper support is seen, however, at the more positive potentials and the electrode did not function as expected after the first sweep to the high positive potentials. Support materials more stable at the high positive potential are required such as nickel and/or alloy nanotubes [155] or aluminium which is typically used in commercial Li ion batteries operating in the potential range of the MnO_2.

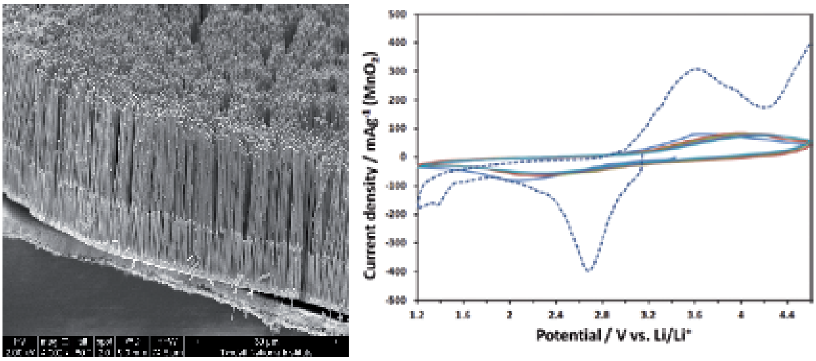


Figure 11. Free standing ramsdellite MnO_2 electrodeposited on Cu nanotubes in an AAO template following removal of the template.

A topochemical method has been described in the synthesis of various nanostructured spinels of $LiMn_2O_4$ with high crystallinity, such as nanowires, nanohorn microspheres and hollow nanospheres, using nanostructured α and γ MnO_2 as precursors. [156]. These nanostructured products are reported mostly as single phase cubic spinel $LiMn_2O_4$ and show higher capacity retention at rates up to 5C. These studies are aimed at improving the structural stability by decreasing the rate of Mn dissolution at elevated temperature. Recently, two reports have been published dealing with improvements in both the rate capabilities and Mn dissolution from metal oxide coatings on spinel $LiMn_2O_4$ nanowires. [157,158] The coated nanowires demonstrate a comparable rate capability to uncoated counterparts but enhanced structural stability in storage at 80°C for 24 hours. Nanotubes of layered $LiCoO_2$, $LiNi_{0.8}Co_{0.2}O_2$ and spinel $LiMn_2O_4$, have been fabricated by AAO template assisted thermal decomposition methods. [159] 1D nanotubes of the open ended lithiated transition metal oxides with uniform shape and sizes and demonstrated gradually decreasing discharge capacities of 170, 150 and 100 mAh/g, respectively, at 100 cycles when cycled at 10 mA/g.

4.3. Nanomaterials for advanced electrode and battery architectures

The majority of commercial batteries are essentially composed of 2D planar films of electrodes and electrolyte. Such an arrangement has been sufficient for energy storage and power delivery to date. However, It is not feasible to have very thick electrodes if the energy stored must be accessed at a high rate which places a limit on the dimensions of the active materials as a result the energy and power capability. For non-aqueous liquid or polymer gel lithium ion batteries this limits the electrode thickness values to approximately 100 μm. Much thinner battery materials (micron scale) are utilised in purely solid state versions such as sputter deposited microbatteries (figure 12). In both cases a limiting factor is the lithium ion diffusion characteristic in the solid state active material.

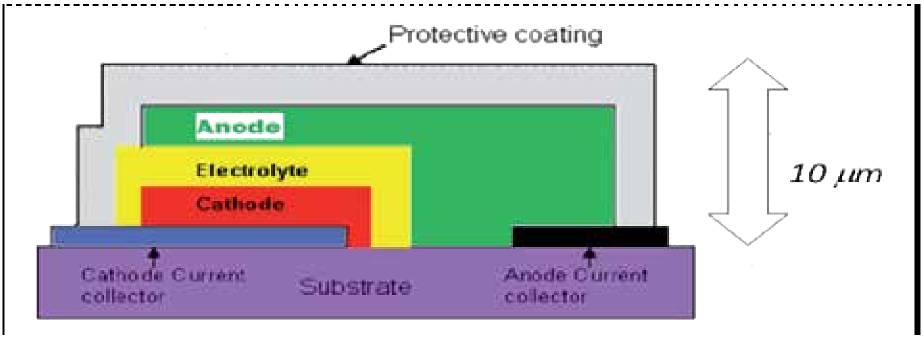


Figure 12. Typical lithium thin film microbattery cross sectional schematic.

Since the mid 1990's [160] work has been progressing on the development of solid state lithium microbatteries appropriate to integration on Si for enhanced microelectronic systems. The development of the LiPON electrolyte has been a significant milestone in microbattery fabrication. [161] Using vacuum deposition techniques planar films of the active components can be deposited and encapsulated leading to thin film energy storage devices that operate with high reproducibility for tens of thousands of cycles. The main issue with such micro energy storage sources is the limited energy that can be stored in micro scale thin films where for example a footprint of 6.5 cm^2 is required to store approximately 4 mWhr. To decrease the footprint the active materials must be structured in high aspect ratio. The most common architectures proposed have been discussed in a detailed review. [112] In general it can be stated that to deliver 4mWh the same energy storage capability in a significantly decreased footprint of 1 mm^2 design which is more appropriate to Si technology the active materials energy capacity must be improved by 3 to 4 times and the materials structured to increase the surface area by 30 times. This is a significant challenge and requires an ability to process active materials at the nanoscale.

2D thin film batteries, figure 13 (a), are limited by the slow transport of Li$^+$ ions and inaccessibility to materials at the back of the plates. On the other hand, 3D or 1D architectured battery materials, figure 13 (b) offer the possibility for short Li$^+$ transport paths to maximise power and energy density if in high aspect ratio. The general strategy for non-planar cell design is to configure electrodes in periodic or aperiodic arrays to obtain the short Li$^+$ transport paths and higher energy density of the cell within the same areal footprint.

The electrochemical performance of 3D batteries based on these architectures mostly depends on the achieved aspect ratios (length/width) and geometries of the electrode. The interdigitated architectures consisting of the separate anode and cathode arrays, is perhaps the most easily envisaged design. [162] The short Li$^+$ transport paths between electrodes and the increased electrode surface area result in a much lower ohmic resist-

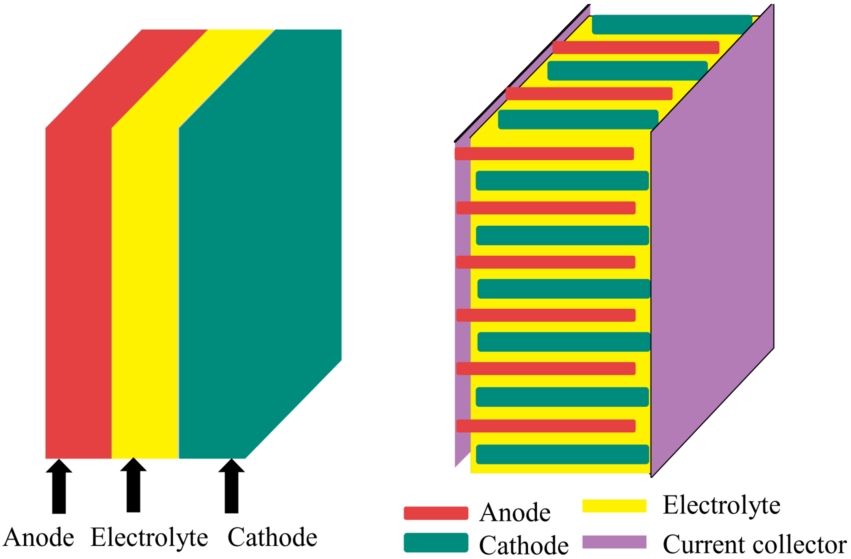


Figure 13. A Schematic illustration of Li-ion microbatteries (a) 2D, parallel thin films and (b) 3D, periodic interdigitated micropillar arrays of anode and cathode are separated by a continuous electrolyte phase.

ance compared to conventional 2D planer counterpart. Initial studies have been carried out on 3D half cell microbatteries consisting of one 3D electrode architecture. One example utilised lithographic methods to amplify the footprint area of 2D thin film battery, essentially by increasing the length (L) of the electrodes to achieve high aspect ratio electrodes. SU-8 negative epoxy based photoresist has been pyrolysed to fabricate high aspect ratio (20:1) C microelectrode arrays following patterning and pyrolysis. [162]

The simplest 3D solid-state battery concept is based on processing of Si substrates to create pillar, figure 14(a) [163] or trench, figure 14(b). [113] In the former case it is difficult to achieve more than 60 micron high features. In addition, sloped sidewalls are desirable to achieve conformal deposition of the active materials using line-of-sight vacuum based deposition processes. Chemical vapour deposition or atomic layer deposition can assist with conformal deposition in higher aspect ratio particularly if thinner layers are required. In this case deep reactive ion etching can create smaller features and thus higher density. Starting from a high surface area Si substrate covered with a current collector, this 3D integrated Li-ion battery concept is based on the successive deposition of diffusion barrier layer of about 70 nm (Ta, TaN or TiN) followed by a high energy dense Si anode thin film of about 50 nm, a solid-state polymer electrolyte LiPON of about 1 μm and a $LiCoO_2$ thin film of about 1 μm, and finally deposition of second current collector. The surface area of the battery can be increased 28-fold compared with a 2D planar thin-film battery with the same foot print. The concept appeals widely for powering Si compatible smart autonomous devices in the integrated chip. The difficulty in this processing is the conformal deposition of successive functional layers in such high aspect ratio features and the fact that a considerable proportion of the device volume is electrically and energetically inactive support material.

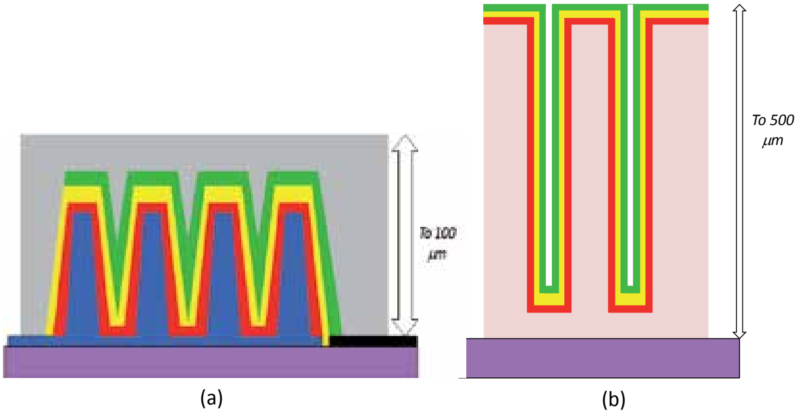


Figure 14. Schematics of (a) additive trench and (b) deep reactive ion etched structures for 3D microbattery fabrication. In these schematics the active materials are represented by the following colour; red = anode, yellow = electrolyte and green = cathode.

The first fully fabricated and characterised 3D microstructured microbatteries were been demonstrated by Nathan et al [164] for concentric 3D tubule electrode arrays. Templates were fabricated in Si or glass and they have achieved very challenging deposition of the successive active material layers in the high aspect ratio arrays and demonstrated outputs 30 times higher than 2D versions. [165] This 3D microbattery technique based on microchannel plates (MCP) uses conformal thin-films of a Ni cathode current collector, a cathode that is mostly composed of MoS_2, a hybrid polymer electrolyte (HPE) and a lithiated MCMB as anode. The cell demonstrated an initial current delivery 2 mA/cm^2, The 3D Li-ion microbatteries retained 60% of their capacity and 100% coulombic efficiency after 200 cycles. An issue with trench or pore etched templates acting as substrates for the energy storage device is the volume they occupy which could in the ideal case be composed of active materials thereby increasing the energy storage density of the device. Colloidal processing of materials has been used to process battery materials. [166] 3D electrodes of vanadium oxide nanorolls (VONR) and mesocarbon microbeads (MCMB) can be used as the cathode and anode for Li-ion battery. [167, 168] Li tests of 3D VONR and MCMB electrodes fabricated by DRIE combined with colloidal process have been carried out in half cell experiments and the results indicate that 3D electrode arrays have significantly higher capacity (45 µAh/mm^2) than the conventional 2D thin film version (6 µAh/mm^2). [159]

Template synthesised inverse opal crystal structures can also be the basis for the fabrication of a 3D interpenetrating electrochemical cells. A fully 3D solid-state interpenetrating Li-ion battery has been demonstrated recently [169] in which a photonic crystal templated C anode is coated by a polymer electrolyte (polyethylene oxide) and the remaining void space is filled with a V_2O_5 ambigel cathode. The fabrication process addresses some issues faced by 3D battery designs, for instance, pre-lithiation of the carbon anode and the addition of electrolyte apparently to increase interfacial contacts. Despite these advances, a major remaining barrier to the fabrication of the rechargeable 3D solid-state interpenetrating microbattery is the low electronic conductivity of the V_2O_5 coupled with the narrow interconnection region between 3D macropores.

Energy storage materials and architectures at the nanoscale is a field of research with many challenges. Some of the design rules and incorporated materials as well as their fabrication strategies have been discussed above. Various 3D architectures and half-cell data has been reported. Optimised full cell versions still requires significant research and development. In particular higher energy density materials with high power performance are required. Matching those capabilities to the available footprint or volume opens up the possibility of a wide range of autonomous microscale devices for future markets. Methods to fabricate and control active materials and architectures at the nanoscale will assist with larger scale versions for improved energy storage in existing markets that require Wh to tens of kWh storage modules.

Acknowledgements

This work has been partially funded by EU Framework 7 project Nanofunction, (Beyond CMOS Nanodevices for Adding Functionalities to CMOS) www.Nanofunction.eu EU ICT Network of Excellence, Grant No.257375 and Enterprise Ireland, Project CFTD07325.

Author details

James F. Rohan, Maksudul Hasan, Sanjay Patil, Declan P. Casey and Tomás Clancy

Tyndall National Institute, University College Cork, Lee Maltings, Cork, Ireland

References

[1] D. Linden and T.B. Reddy, in *Handbook of Batteries*, 3rd eds., (2002).

[2] H. Ikeda, T. Saito and M. Tamura, in *Proc. Manganese Dioxide Symp.* (eds. A. Kozawa and H.R. Brodd) (Cleveland, OH), 1, (1975).

[3] M.S. Whittingham, *Science*, 192, (1976) 1226.

[4] M.S. Whittingham, Chalcogenide battery, US patent 4009052.

[5] B.M.L. Rao, R.W. Francis and H.A. Christopher, *J. Electrochem. Soc.*, 124, (1977) 1490.

[6] D.W. Murphy, F.J. DiSalvo, J.N. Carides and J.V. Waszczak, *Mat. Res. Bull.*, 13, (1978) 1395.

[7] M. Lazzari and B. Scrosati, *J. Electrochem. Soc.*, 127, (1980) 773.

[8] J. Broadhead and A.D. Butherus, US patent 3791867.

[9] J. Broadhead, F.J. DiSalvo and F.A. Trumbore, US patent 3864167.

[10] K. Mizushima, P.C. Jones, P.J. Wiseman and J.B. Goodenough, *Mat. Res. Bull.*, 15, (1980) 783.

[11] M.M. Thackeray, W.I.F David, P.G. Bruce and J.B. Goodenough, *Mat. Res. Bull.*, 18, (1983) 461.

[12] S. Basu, US patent 4423125.

[13] M. Mohri, N. Yanagisawa, Y. Tajima, H. Tanaka, T. Mitate, S. Nakajima, Y. Yoshida, Y. Yoshimoto, T. Suzuki, and H. Wada, *J. Power Sources*, 26, (1989) 545.

[14] T. Nagaura and K. Tozawa, *Prog. Batteries Solar Cells*, 9, (1990) 209.

[15] M. Armand, J.M. Chabagno and M.J. Duclot, in *Fast Ion transport in Solids Electrodes and Electrolytes* (eds. P. Vashishta, J.N. Mundy and G.K. Shenoy), North-Holland, Amsterdam, (1979) 131.

[16] K.M. Abraham and M. Alamgir, *J. Electrochem. Soc.*, 137, (1990) 1657.

[17] K.M. Abraham and M. Alamgir, *J. Power Sources*, 44, (1993) 195.

[18] J.M. Tarascon, A.S. Gozdz, C. Schmutz, F. Shokoohi and P.C. Warren, *Solid State Ionics*, 86-88, (1996) 49.

[19] Z.X. Shu, R.S. McMillan and J.J. Murray, *J. Electrochem. Soc.*, 140, (1993) 922.

[20] J.M. Tarascon and D. Guyomard, *Electrochim. Acta*, 38, (1993) 1221.

[21] T. Ohzuku, Y. Iwakoshi and K. Swai, *J. Electrochem. Soc.*, 140, (1993) 2490.

[22] J.R. Dahn, A.K. Sleigh, H. Shi, J.N. Reimers, Q. Zhong and B.M. Way, *Electrochim. Acta*, 38, (1993) 1179.

[23] N. Sonobe, M. Ishikawa and T. Iwasaki, in *Proceedings of the Abstracts of 35th Battery Symposium*, Nagoya, Japan, (1994) 47.

[24] T. Takamura, K. Sumiya, J. Suzuki, C. Yamada and K. Sekine, *J. Power Sources*, 81-82, (1999) 368.

[25] M. Winter and J.O. Besenhard, *Electrochim Acta*, 45, (1999) 31.

[26] Anani, S. Crouch-Baker and R.A. Huggins, *J. Electrochem. Soc.*, 134, (1987) 3098.

[27] O. Mao, R.A. Dunlap and J.R. Dahn, *J. Electrochem. Soc.*, 146, (1999) 405.

[28] J.J. Zhang and Y.Y. Xia, *J. Electrochem. Soc.*, 153, (2006) A1466.

[29] L. Wang, S. Kitamura, K. Obata, S. Tanase and T. Sakai, *J. Power Sources*, 141, (2005) 286.

[30] H. Mukaibo, T. Momma and T. Osaka, *J. Power Sources*, 146, (2005) 457.

[31] K.D. Kepler, J.T. Vaughey and M.M. Thackeray, *Electrochem. Solid-State Lett.*, 2, (1999) 307.

[32] H. Li, L. Shi, Q. Wang, L. Chen and X. Huang, *Solid State Ionics*, 148, (2002) 247.

[33] Y. Idota, T. Kabuto, A. Matsufuji, Y. Maekawa and T. Miyasaki, *Science*, 276, (1997) 1395.

[34] O. Mao, and J.R. Dahn, *J. Electrochem. Soc.*, 146, (1999) 423.

[35] L.Y. Beaulieu and J.R. Dahn, *J. Electrochem. Soc.*, 147, (2000) 3237.

[36] G.X. Wang, J.H. Ahn, J. Yao, S. Bewlay and H.K. Liu, *Electrochem. Commun.*, 6, (2004) 689.

[37] H. Kim, and J. Cho, *J. Electrochem. Soc.*, 154, (2007) A462.

[38] N. Jayaprakash, N. Kalaiselvi, C. H. Doh, *J Appl. Electrochem.*, 37 (2007) 567

[39] Y. Wang, F. Su, J.Y. Lee and X.S. Zhao, *Chem. Mater.*, 18, (2006) 1347.

[40] X.W. Lou, Y. Wang, C. Yuan. J.Y. Lee and L.A. Archer, *Adv. Mater.*, 18, (2006) 2325.

[41] H. Kim and J. Cho, *Chem. Mater.*, 20, (2008) 1679.

[42] S.W. Kim, M. Kim, W.Y. Lee and T. Hyeon, *J. Am. Chem. Soc.*, 124, (2002) 7642.

[43] M. Yang, J. Ma, C. Zhang, Z. Yang and Y.Lu, *Angew. Chem. Int. Ed.*, 44, (2005) 6727.

[44] J. Gao, B. Zhang, X. Zhang and B. Xu, *Angew. Chem. Int. Ed.*, 45, (2006) 1220.

[45] Y. Wang, J.Y. Lee and H.C. Zeng, *Chem. Mater.*, 17, (2005) 3899.

[46] Y. Wang, H.C. Zeng and J.Y. Lee, *Adv. Mater.*, 18, (2006) 645.

[47] W-M. Zhang, J-S. Hu, Y-G. Guo, S-F. Zheng, L-S. Zhong, W-G. Song and L-J. Wan, *Adv. Mater.*, 20, (2008) 1160.

[48] T. Ohzuku and A. Ueda, *Solid State Ionics*, 69, (1994) 201.

[49] Y. Gao, M.V. Yakovleva and W.B. Ebner, *Electrochem. Solid State Lett.*, 1, (1998) 117.

[50] C. Pouillerie, F. Perton, P. Biensan, J.P. Peres, M. Broussely and C. Delmas, *J. Power Sources*, 96, (2001) 293.

[51] C. Pouillerie, L. Croguennec and C. Delmas, *Solid State Ionics*, 132, (2000) 15.

[52] I. Nakai and T. Nakagome, *Electrochem. Solid State Lett.*, 1, (1998) 259.

[53] M. Brouselly, *Lithium Battery Discussion*, Bordeaux-Archachon (2001).

[54] R. Chen and M.S. Whittingham, *J. Electrochem. Soc.*, 144, (1997) L64.

[55] F. Capitaine, P. Gravereau and C. Delmas, *Solid State Ionics*, 89, (1996) 197.

[56] R. Stoyanova, E. Zhecheva and L. Zarkova, *Solid State Ionics*, 73, (1994) 233.

[57] K. Numata and S. Yamanaka, *Solid State Ionics*, 118, (1999) 117.

[58] M.E Spahr, P. Novak, B. Schnyder, O.Haas and R. Nesper, *J. Electrochem. Soc.*, 145, (1998) 1113.

[59] T. Ohzuku and Y. Makimura, *Chem. Lett.*, 8, (2001) 744.

[60] S. Yang, Y. Song, K. Ngala, P.Y. Zavalij and M.S. Whittingham, *J. Power Sources*, 119, (2003) 239.

[61] Z. Wang, Y. Sun, L. Chen and X. Huang, *J. Electrochem. Soc.*, 151, (2004) A914.

[62] Y.W. Tsai, J.F. Lee, D.G. Liu and B. Hwamg, *J. Mater. Chem.*, 14, (2004) 958.

[63] S.H. Park, C.S. Yoon, S.G. Kang, H-S. Kim, S-I. Moon and Y-K. Sun, *Electrochim. Acta*, 49, (2004) 557.

[64] N. Yabuuchi and T. Ohzuku, *J. Power Sources*, 171, (2003) 119.

[65] J.M. Kim and H.T. Chung, *Electrochim. Acta*, 49, (2004) 937.

[66] S. Jouanneau, D.D. Macneil, Z. Lu, S.D. Beattie, G. Murphy and J.R. Dahn, *J. Electrochem. Soc.*, 150, (2003) A1299.

[67] N. Ravet, Y. Chouinard, J.F. Magnan, S. Besner, M. Gauthier, and M. Armand, *J. Power Sources.*, 97-98, (2001) 503.

[68] H. Huang, S.-C. Yin and J.F. Nazar, *Electrochem. Solid-State Lett.*, 4, (2001) A170.

[69] J.-M. Tarascon and M. Armand, *Nature*, 541, (2008) 652.

[70] R. Dominko, M.Bele, M. Gaberscek, M. Remskar, D. Hanzel, S. Pejovnik and J. Jamnik, *J. Electrochem. Soc.*, 152, (2005) A607.

[71] R. Dominko, M.Bele, M. Gaberscek, M. Remskar, D. Hanzel, J.M. Goupil, S. Pejovnik and J. Jamnik, *Solid State Ionics*, 153, (2006) 274.

[72] M. Gaberscek and J. Jamnik, *Solid State Ionics*, 177, (2006) 2647.

[73] Y.-H Huang, K.-S Park and J.B. Goodenough, *J. Electrochem. Soc.*, 153, (2006) A2282.

[74] Y.-S. Hu, Y.-G. Guo, R. Dominko, M. Gaberscek, J. Jamnik and J. Maier, *Adv. Mater.*, 19 (2007) 1963.

[75] M.S. Whittingham, *J. Electrochem. Soc.*, 123, (1976) 315.

[76] F.W. Dampier, *J. Electrochem. Soc.*, 121, (1974) 656.

[77] R.J. Gummow, A.D. Kock and M.M. Thackeray, *Solid State Ionics*, 69, (1994) 59.

[78] C. Delmas, H. C. Auradou, J.M. Cocciantelli, M. Ménétrier and J.P. Doumerc, *Solid State Ionics*, 69, (1994) 257.

[79] D.W. Murphy, P.A Christian, F.J. Disalvo and J.N. Carides, *J. Electrochem. Soc.*, 126, (1979) 497.

[80] J.O. Besenhard and R. Schöllhorn, *J. Power Sources*, 1, (1976) 267.

[81] K. Nassau and D.W. Murphy, *J. Non-Cryst. Solids*, 44, (1981) 297.

[82] P.S. Herle, B. Ellis, N. Coombs and L.F. Nazar, *Nat. Mater.*, 1 (2002) 123.

[83] S.-Y. Chung, J.T. Bloking and Y.-M. Chiang, *Nat. Mater.*, 1 (2002) 123.

[84] Q. Wang, J. Sun, X. Yao and C. Chen, *Thermochimica Acta*, 437, (2005) 12.

[85] T. Kawamura, A. Kimura, M. Egashira, S. Okada and J.I. Yamaki, *J. Power Sources*, 104, (2002) 260.

[86] H. Yang, G. V. Zhuang and P. N. Ross Jr., *J. Power Sources*, 161, (2006) 573.

[87] S.I. Lee, U.H. Jung, Y.S. Kim, M.H. Kim, D.J. Ahn and H.S. Chun, *Korean J. Chem. Eng.*, 19, (2002) 638.

[88] M. Morita, M. Ishikawa and Y. Matsuda, in *Lithium Ion Batteries* "Organic electrolytes for rechargeable lithium ion batteries" (eds. M. Wakihara, O. Yamamoto), Kodansha, Tokyo, Wiley/VCH, Weinheim, (1998) 156.

[89] D. Aurbach, A. Zaban, Y. Ein-Eli, I. Weissman, O. Chusid, B. Markovsky, M. Levi, E. Levi, A. Schechter and E. Granot, *J. Power Sources*, 68, (1997) 91.

[90] S.E Sloop, J.B. Kerr and K. Kinoshita, *J. Power Sources*, 119-121, (2003) 330.

[91] P. Liu and H. Wu, *J. Power Sources*, 56, (1995) 81.

[92] M. Arakawa and J. Yamaki, *J. Electroanal. Chem.*, 219, (1987) 273.

[93] M. Alamgir and K.M. Abraham, in *Lithium Batteries* "New Materials, Developments and Perspectives" (ed. G. Pistoia), Amsterdam, Elsevier, (1994) p 93.

[94] G. Feuillade and P. Perche, *J. Appl. Electrochem.*, 5, (1975) 63.

[95] S. Chintapalli and R. Frech, *Solid State Ionics*, 86-88, (1996) 341.

[96] E. Tsuchida, H. Ohno and K. Tsunemi, *Electrochimica Acta*, 28, (1983) 591.

[97] N.S. Mohamed and A.K. Arof, *J. Power Sources*, 132, (2004) 229.

[98] M. Watanabe, M. Kanba, K. Nagaoka and I. Shinohara, *J. Appl. Polym. Sci.*, 27, (1982) 4191.

[99] G.B. Appetecchi and B. Scrosati, *Electrochimica Acta*, 43, (1998) 1105.

[100] G.B. Appetecchi, F. Croce and B. Scrosati, *Electrochimica Acta*, 40, (1995) 991.

[101] J. Vondrak, M. Sedlarikova, J. Velicka, B. Klapste, V. Novak and J. Reoter, *Electrochimica Acta*, 46, (2001) 2047.

[102] C. Capiglia, Y. Saito, H. Kataoka, T. Kodama, E. Quartarone and P. Mustarelli, *Solid State Ionics*, 131, (2001) 291.

[103] D. Saikia and A. Kumar, *Electrochimica Acta*, 49, (2004) 2581.

[104] F. Croce, F. Gerace, G. Duatzemberg, S. Passerini, G.B. Appetecchi and B. Scrosati, *Electrochimica Acta*, 39, (1994) 2187.

[105] Z. Jiang, B. Carroll and K.M. Abraham, *Electrochimica Acta*, 42, (1997) 2667.

[106] A.S. Gozdz, C. Schmutz and J.M. Tarascon, *US Patent No* 5296318

[107] M.M.E. Jacob, E. Hackett and E.P. Giannelis, *J. Mater. Chem.*, 13, (2003) 1.

[108] S. Abbrent, S.H. Chung, S.G. Greenbaum, J. Muthu and E.P. Giannelis, *Electrochimica Acta*, 48, (2003) 2113.

[109] M. Wachtler, D. Ostrovskii, P. Jacobsson and B. Scrosati, *Electrochimica Acta*, 50, (2004) 357.

[110] V. Gentili, S. Panero, P. Reale and B. Scrosati, *J. Power Sources*, 170, (2007) 185.

[111] Y. Wang, V. Schmidt, S. Senz and U. Gosele, *Nat. Nanotechnol.*, 1, (2006) 189.

[112] J.W. Long, B. Dunn, D.R. Rolison, H.S. White, Chem. Rev. 104 (2004) 4463

[113] L. Baggetto, R.A.H. Niessen, F. Roozeboom, P.H.L. Notten, Adv. Funct. Mater., 18 (2008).1057

[114] C. R. Sides, C. R. Martin, *Nanomaterials in Li-Ion Battery Electrode Design, Modern Aspects of Electrochemistry, no. 40*, R. E. White, Editor, p. 75, Springer, New York (2007)

[115] C. R. Martin, Adv. Mater. 3 (1991) 457

[116] J.C. Hulteen and C.R. Martin, *J. Mater. Chem.*, 7, (1997) 1075.

[117] S. A. Sapp, B.B. Lakshmi, C. R. Martin, *Adv. Mater*. 12 (1999) 402

[118] H. Masuda, K. Fukuda, *Science* 268 (1995) 1466

[119] S. Shingubara, O. Okino, Y. Sayama, H. Sakaue, T. Takahagi, *Solid-State Electronics* 43 (1999) 1143

[120] D. Crouse, Y.H. Lo, A.E. Miller, M. Crouse, *Appl. Phys. Lett*. 76 (2000) 49

[121] N.V. Myung, J. Lim, J-P. Fleurial, M. Yun, W. West, D. Choi, *Nanotechnology*, 15 (2004) 833

[122] J-H. Kim, S. Khanal, M. Islam, A. Khatri, D. Choi, *Electrochem. Commun*, 10 (2008) 1688

[123] F. Cheng, Z. Tao, J. Liang, J. Chen, *Chem. Mater*. 20 (2008) 667

[124] A.M. Morales and C.M. Lieber, *Science*, 279, (1998) 208.

[125] C.K. Chan, X.F. Zhang and Y. Cui, *Nano Lett.*, 8, (2008) 307.

[126] N. Zhao, G. Wang, Y. Huang, B. Wang, B. Yao and Y. Wu, *Chem. Mater.*, 20, (2008) 2612.

[127] A.S. Arico, P. Bruce, B. Scrosati, J.-M. Tarascon, W.V. Schalkwijk, *Nat. Mater.*, 4, (2005) 366.

[128] C.K. Chan, H. Peng, R.D. Twesten, K. Jarausch, X.F. Zhang and Y. Cui, *Nano Lett.*, 7, (2007) 490.

[129] J. O. Besenhard, J. Yang and M. Winter, *J. Power Sources*, 68, (1997) 87.

[130] J.F. Rohan, M. Hasan and N. Holubowitch, *Electrochim. Acta,* 56 (2011) 9537

[131] D. Larcher, L.Y. Beaulieu, D.D. MacNeil, J.R. Dahn*, J. Electrochem. Soc.,* 147 (2000) 1658.

[132] L. Fransson, E. Nordstrom, K. Edstrom, L. Haggstrom, J.T. Vaughey, M.M. Thackeray, *J. Electrochem. Soc*. 149 (2002) A736

[133] A.H. Whitehead, J.M. Elliott, J.R. Owen, *J. Power Sources* 81–82 (1999) 33

[134] J. Graetz, C.C. Ahn, R. Yazami and B. Fultz, *J. Electrochem. Soc.*, 151, (2004) A698.

[135] Y. Wang, X. Jiang and Y. Xia, *J. Am. Chem. Soc.*, 125, (2003) 16176.

[136] N. Ramgir, I. Mulla and K. Vijayamohanan, *J. Phys. Chem. B*, 108, (2004) 14815.

[137] W. Yu, X. Li, X. Gao and F. Wu, *J. Phys. Chem. B*, 109, (2005) 17078.

[138] Y. Wang, J.Y. Lee and H.C. Zeng, *Chem. Mater.*, 17, (2005) 3899.

[139] Y. Wang, H.C. Zeng and J.Y. Lee, *Adv. Mater.*, 18, (2006) 645.

[140] P.L. Taberna, S. Mitra, P. Poizot, P. Simon and J.-M. Tarascon, *Nature*, 5, (2006) 567.

[141] Y. Li, B. Tan and Y. Wu, *Nano Lett.*, 8, (2008) 265.

[142] T. Chowdhury, D.P. Casey and J.F. Rohan, *Electrochem. Commun.*, 11 (2009) 1203

[143] M. Hasan, T. Chowdhury and J.F. Rohan, *J. Electrochem. Soc.*, 157 (2010) A682

[144] C.K. Chan, H. Peng, G. Liu, K.M. Wrath, X.F. Zhang, R.A. Huggins and Y. Cui, *Nature*, 3, (2008) 31

[145] C.K. Chan *J. Power Sources* 189 (2009) 34–39

[146] L-F. Cui, R. Ruffo, C.K. Chan, H. Peng, Y. Cui, *Nano Lett.,* 9 (2009) 1

[147] H. Kim and J. Cho, *Nano Lett.*, 8, (2008) 3688.

[148] C.J. Patrissi and C.R. Martin, *J. Electrochem. Soc.*, 146, (1999) 3176.

[149] C.R. Sides and C.R. Martin, *Adv. Mater.*, 17, (2005) 125.

[150] K. Takahashi, S. J. Limmer, Y. Wang and C.Z. Cao, *J. Phys. Chem. B*, 108, (2004) 9795.

[151] K. Takahashi, Y. Wang and C.Z. Cao, *Appl. Phys. Lett.*, 86, (2005) O53102.

[152] M.E. Spahr, P. Bitterli, R. Nesper, M. Müller, F. Krumeich and H.U. Nissen, *Angew. Chem. Int. Ed.*, 37, (1998) 1263.

[153] Y. Wang, K. Takahashi, H. Shang and G.Z. Cao, *J. Phys. Chem. B*, 109, (2005) 3085.

[154] W.C. West, N.V. Myung, J.F. Whitacre and B.V. Ratnakumar, *J. Power Sources*, 126, (2004) 203.

[155] J.F. Rohan, D.P. Casey, B. Ahern, F.M.F. Rhen, S. Roy, D. Fleming and S.E. Lawrence, *Electrochem. Commun*, 2008 10 1419

[156] J-Y. Luo, H-M. Xiong and Y-Y. Xia, *J. Phys. Chem. C*, 112, (2008) 12051.

[157] J. Cho, *J. Mater. Chem.*, 18, (2008) 2257.

[158] S.H. Lim and J. Cho, *Electrochem. Commun.*, 10, (2008) 1478.

[159] X. Li, F. Cheng, B. Guo and J. Chen, *J. Phys. Chem. B*, 109, (2005) 14017.

[160] S.D. Jones, J.R. Akridge, *J. Power Sources* 54 (1995) 63

[161] J.B. Bates, N.J. Dudney, D.C. Lubben, G.R. Gruzalski, B.S. Kwak, X. Yu, R.A. Zuhr, J. Power Sources 54 (1995) 58

[162] C. Wang, L. Taherabadi, G. Jia, M. Madou, Y. Yeh and B. Dunn, *Electrochem. Solid-State Lett.* 7, (2004) A435.

[163] A.V. Jeyaseelan and J.F. Rohan, *Appl. Surf. Science* 256S (2009) S61

[164] M. Nathan, D. Golodnitsky, V. Yufit, E. Strauss, T. Ripenbein, I. Shechtman, S. Menkin and E. Peled, *J. MEMS*, 14, (2005) 879.

[165] D. Golodnitsky, V. Yufit, M. Nathan, I. Shechtman, T. Ripenbein, E. Strauss, S. Menkin and E. Peled, *J. Power Sources*, 153, (2006) 281.

[166] J.A. Lewis, *J. Amer. Ceram. Soc.*, 83, (2000) 2341.

[167] D. Sun, C.W. Kwon, G. Baure, E. Richman, J. MacLean, B. Dunn and S.H. Tolbert, *Adv. Funct. Mater.*, 14, (2004) 1197.

[168] J. Yao, G.X. Wang, J.-H. Ahn, H.K. Liu and S.X. Dou, *J. Power Sources*, 114, (2003) 292.

[169] N.S. Ergang, M.A. Fierke, Z, Wang, W.H. Smyrl and A. Stein, *J. Electrochem. Soc.*, 154, (2007) A1135.

Chapter 7

Minimum Energy of Computing, Fundamental Considerations

Victor Zhirnov, Ralph Cavin and Luca Gammaitoni

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57346

1. Introduction

Energy consumption during computation has become a matter of strategic importance for modern ICT and its impact on future society. In this chapter we review some of the basic principles governing energy consumption in ICT and discuss future perspectives toward more efficient computers.

In the last forty years the progress of the semiconductor industry has been driven by its ability to cost-effectively scale down the size of the CMOS-FET [1] switches, the building block of present computing devices, and this has provided continuing increases in computing capability. However, this has been accompanied by a continuing increase in energy consumption and heat generation up to a point where the power dissipated in heat during computation has become a serious limitation [2, 3]. The energetics issues of current and future computation raises a question of the ultimate *energy efficiency* of computation that is reminiscent of the Carnot limit for the efficiency for heat engines [4]. It should be noted that the entire discipline of thermodynamics emerged from the practical need to increase the efficiency of utility heat engines. Innovations in steam engines and internal combustion engines have been driven by the need to more closely approach the ideal limit of a Carnot engine. Today, approximately 200 years after the work of Carnot, the problem of understanding the efficiency of generalized energy generators remains, although today the object of interest not only includes large power plants but also small scale devices and systems for information processing. In fact, one can view information processor as a *computing engine* that transforms incoming energy flow into useful work and also produces some heat [5].

Interesting insights on the energy efficiency of binary elements were obtained in pioneering studies by John von Neumann and subsequently by Charles H. Bennet and Ralph Landauer in the last century. It has been shown that information processing is intimately related to energy management ("information is physical"). Specifically, Bennet and Landauer have shown that there exists a minimum amount of energy required to perform any irreversible computation. The ultimate limit on the minimum energy per switching is set at $k_B T\ln2$ (approximately 10^{-21} J at room temperature) [6, 7] called the Shannon-von Neumann-Landauer (SNL) limit [8]. As Landauer argued, this minimum amount cannot be reduced to zero if some information is discarded (erased) during the computation process. The reason is directly linked to thermodynamics: erasing information decreases the overall entropy of the system and this cannot be done without dissipating heat of at least k_BT ln2 Joule per bit erased [7, 8].

While the physical limits of individual binary elements (switches) have been explored to a significant depth (many questions remain open however), currently, there are no theoretical results available that characterize the maximum computational efficiency of a computing systems as a whole; for example, in the spirit of the bound on efficiency of heat engines obtained by Carnot. A full understanding of possible limits to computational performance similar to the Carnot efficiency limit for heat engines would be extremely important both from theoretical point of view and could guide the design of future extremely energy-efficient computational engines. As an example, the Nanoelectronics Research Initiative was launched in 2005 [3], funded by a US-based consortium of semiconductor companies, and federal and state governments, to address a grand challenge to understand the fundamental energy limits of the physics of both binary elements (logic and memory) and computing systems. Before exploring the basic principles that govern minimum energy dissipation in devices and computation, it is appropriate to briefly review the state of the art for present computers.

2. Energy dissipation in present computers — A field survey

The four main information processing functions in modern electronic ICT systems are: computation, communication, storage, and display, as shown in Fig. 1 [9]. In the U. S. alone

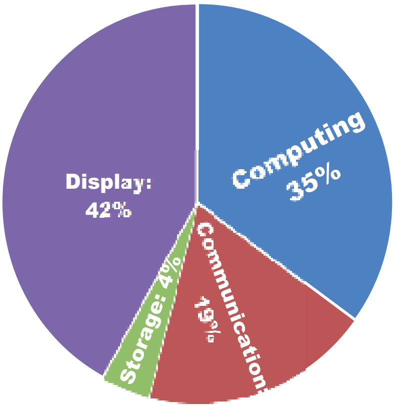


Figure 1. Energy consumption by four primary information processing functions in modern electronic ICT systems [9].

these constitutes about 290 TWh/year of electrical power, costing ~$30 Billion/year and the amount of electrical energy consumed by ICT is expected to continue to grow. It is instructive to review the sources of energy consumption in state-of-the-art ICT systems, since this may offer insights for possible directions to improve their energy efficiency.

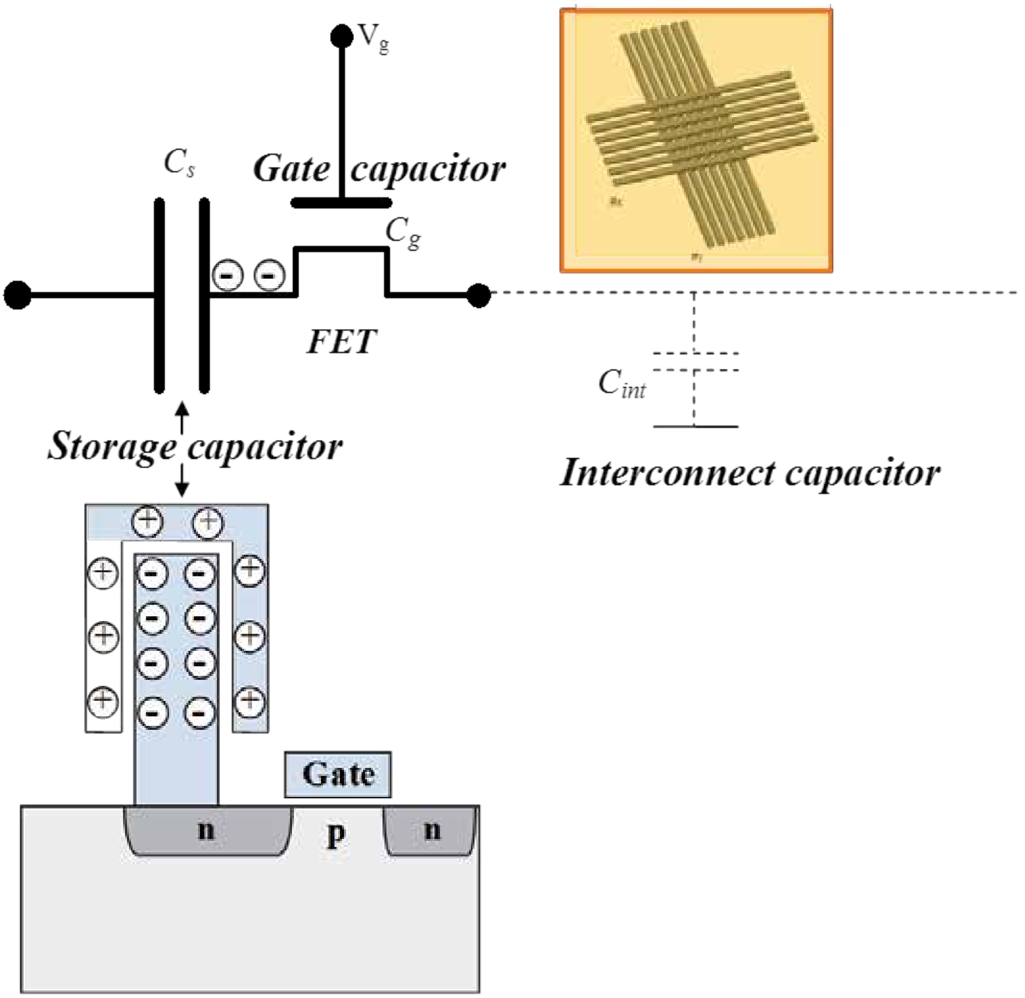


Figure 2. Device and circuit capacitance as a central concept of microelectronics.

2.1. Logic and memory devices

The main source of energy consumption in electronics is charging and discharging of electrical capacitances, which are present in all electronic devices. To illustrate this, an example of a dynamic random-access memory (DRAM) where several distinct capacitors are present, is shown in Fig.2 including a storage capacitor C_s, the gate capacitor C_g of the field effect transistor

(FET), and interconnect capacitor C_{int} formed by the wires used to connect individual memory cells in an X-Y array.

Energy dissipation by charging a capacitor is a central concept of microelectronics as operation of all electronic devices involves charging/discharging corresponding capacitors. When a capacitor is charged from a constant voltage power supply, energy is dissipated, i.e. converted into heat. Consider a typical model circuit consisting of a capacitor *C* in series with a resistor *R* (Fig. 3). Suppose a constant voltage of magnitude *V* is applied to the circuit at *t*=0 and electrical charge flows to the capacitor. The charging of the capacitor is characterized by a time-dependent voltage drops both on the resistor and the capacitor:

$$V_R(t) = V\exp\left(-\frac{t}{RC}\right) \tag{1}$$

$$V_C(t) = V\left(1-\exp\left(-\frac{t}{RC}\right)\right) \tag{2}$$

The energy dissipated in the resistor *R* during charging is:

$$E_R = \int_0^\infty \frac{V_R^2(t)}{R}\,dt = \frac{V}{R}\int_0^\infty [\exp(-\frac{t}{RC})]^2\,dt = \frac{CV^2}{2} \tag{3}$$

Note that the energy dissipated in the resistor is independent of the resistance value *R*. As result of the charging process, the capacitor stores the energy $E_C = ½CV^2$, and thus the total energy required for constant charging voltage (the switching energy) is:

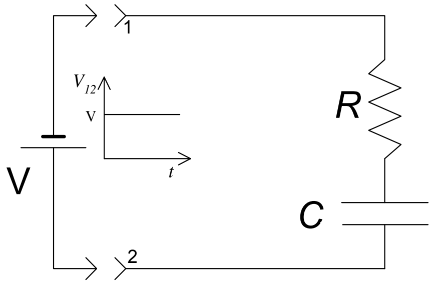


Figure 3. Generic *RC* circuit.

$$E_{sw} = CV^2 \tag{4}$$

Now capacitance is related to a linear dimension, *L*:

$$C \sim \varepsilon_0 L \tag{5}$$

If binary switching (i.e. capacitor charging and discharging) occurs with a frequency *f*, the operational power is:

$$P = \alpha E_{sw} f = \alpha C V^2 f \tag{6}$$

Where α is the activity factor, α=1, for a square-wave switching and α<0.5 in typical digital circuits. In a circuit with a large number of transistors, N_{tr}, the total switching energy is $E_{\mathrm{sw}} = N_{\mathrm{tr}}\ E_{\mathrm{sw1}}$, where E_{sw1} is the switching energy of an individual transistor. Note that (4) and (6) refer to *dynamic* switching energy and power, directly related to the ON/OFF switching. There is also a parasitic leakage power component that will be discussed in the following.

According to (4), on the level of elementary operations (e.g. binary switching), the control space is limited by only two parameters – operating voltage and device and capacitance (devices and interconnects), and both these parameters have been considerably reduced during past 40 years as result of *scaling* – the continuing decrease of the device critical dimension, *F*, from tenths of micrometers to only a few nanometers. One immediate implication of scaling is a linear decrease of device capacitance (e.g. the FET gate capacitance, C_g) with F: C_g=k_1F. If voltage can also be linearly scaled with the device size with some coefficient of proportionality k_2, i.e. V=k_2F then the FET switching energy (CV^2) decreases as the cube of the device dimensions:

$$E_{sw1} = C_g V^2 = k_1 F \cdot \left(k_2 F\right)^2 = k_1 k_2 F^3 \tag{7}$$

Switching energies of individual transistors for several generations (1994-2011) of microprocessor units (MPU) are shown in Table 1 and Fig. 4. The data points for this 'bottom-up' approach were taken from several editions of the International Technology Roadmap for Semiconductors (ITRS) [10]. Note that the data points are approximated by a nearly cubic power function with strong correlation (the determination coefficient R^2 = 0.97), which is consistent with (7).

Now, it is instructive to compare the 'bottom-up' number with a 'top-down' average energy per transistor calculated from total power dissipation in practical microprocessors. From (6):

$$\left\langle E_{sw1} \right\rangle_{MPU} = \frac{P}{\alpha f N_{tr}} \tag{8}$$

The 'top-down' data points plot in Fig. 4 were obtained using (8), data from Table 1 and assuming the activity factor α=0.25.

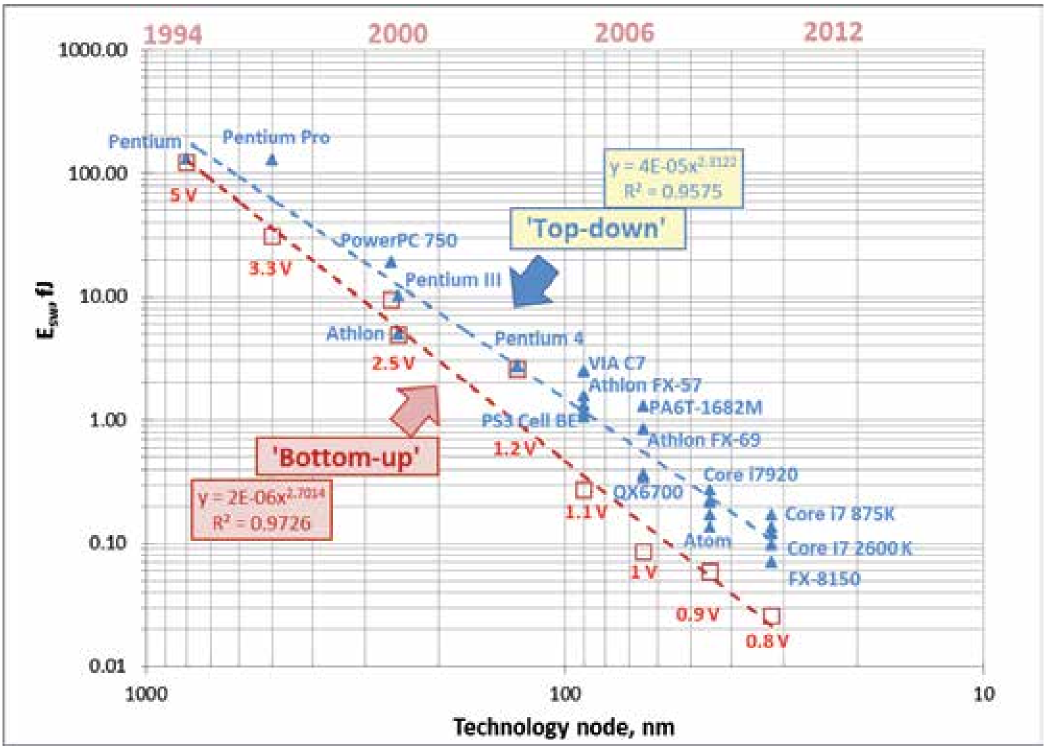


Figure 4. Switching energies of individual transistors in several generations (1994-2011) of microprocessors, calculated using 'bottom-up and 'top-down' approaches

Comparison of the 'top-down' and 'bottom-up' lines shows a clear divergence of the two as scaling continues. For a better visualization, a plot in Fig. 5 shows the ratio of the individual transistor dynamic switching energy ('bottom-up') to the average energy per transistor in MPU ('top down'):$\frac{E_{sw1}}{\langle E_{sw1}\rangle_{MPU}}$.

While for larger scale devices (e.g. ~ 1μm), the switching energy of transistor is a determining factor in the total chip energy consumption, for sub-100 nm technology nodes, the fraction of the transistor dynamic energy in the energy balance decreases. Indeed, transistor dynamic energy consumption constitutes ~20-30% of the total energy in modern microprocessors (22-65 nm node), and is expected to further decrease for 10nm devices and below. This trend is driven by increased dissipation in interconnects and off-state leakage losses.

As follows from the above, wires connecting binary switches, constitute a significant (and often dominant) portion of the energy consumption in ITC, and suggests a Carnot-type efficiency limit for computational engines.

Processor	Year	F, nm	IPS	f, MHz	N_{tr}	P, W	'bottom-up' E_{sw}, fJ	'top-down' E_{sw}, fJ
Intel Pentium	1994	800	1.88E+08	100	3.10E+06	10.1	122.72	130.32
Intel Pentium Pro	1996	500	5.41E+08	200	5.50E+06	35.0	31.32	127.27
PowerPC 750	1997	260	5.25E+08	233	6.35E+06	7.0	9.40	18.92
Intel Pentium III	1999	250	2.05E+09	600	9.50E+06	7.0	4.93	4.91
AMD Athlon	2000	250	3.56E+09	1200	2.20E+07	65.7	4.93	9.95
AMD Athlon XP 2500+	2003	130	7.53E+09	1830	5.43E+07	68.0	2.59	2.74
Pentium 4 Extreme Edition	2003	130	9.73E+09	3200	5.50E+07	115.0	2.59	2.61
VIA C7	2005	90	1.80E+09	1300	2.50E+07	20.0	0.27	2.46
AMD Athlon FX-57	2005	90	1.20E+10	2800	1.14E+08	104.0	0.27	1.30
AMD Athlon 64 3800+ X2 (Dual core)	2005	90	1.46E+10	2000	1.54E+08	89.0	0.27	1.16
Xbox360 IBM "Xenon" (Triple core)	2005	90	1.92E+10	3200	1.65E+08	203.0	0.27	1.54
PS3 Cell BE (PPE only)	2006	90	1.02E+10	3200	2.41E+08	200.0	0.27	1.04
AMD Athlon FX-60 (Dual core)	2006	65	1.89E+10	2600	2.33E+08	125.0	0.09	0.82
Intel Core 2 Extreme X6800 (Dual core)	2006	65	2.71E+10	2930	2.91E+08	75.0	0.09	0.35
Intel Core 2 Extreme QX6700 (Quad core)	2006	65	4.92E+10	2660	5.82E+08	130.0	0.09	0.34
P.A. Semi PA6T-1682M	2007	65	8.80E+09	2000	1.10E+07	7.0	0.09	1.27
Intel Core 2 Extreme QX9770	2008	45	5.95E+10	3200	8.00E+08	136.0	0.06	0.21
Intel Core i7 920	2008	45	8.23E+10	2660	7.31E+08	130.0	0.06	0.27
Intel Atom N270	2008	45	3.85E+09	1600	4.70E+07	2.5	0.06	0.13
AMD Phenom II X4 940	2009	45	4.28E+10	3000	7.58E+08	125.0	0.06	0.22
AMD Phenom II X6 1100T Thuban	2010	45	7.84E+10	3300	9.04E+08	125.0	0.06	0.17
Intel Core i7 Extreme Edition 980X	2010	32	1.48E+11	3330	1.17E+09	130.0	0.03	0.13
Intel Core i7 2600K	2011	32	1.28E+11	3400	1.16E+09	95.0	0.03	0.10
AMD E-350	2011	32	1.00E+10	1600	3.80E+08	18.0	0.03	0.12
Intel Core i7 875K	2011	32	9.21E+10	2930	7.74E+08	95.0	0.03	0.17
AMD FX-8150	2011	32	1.09E+11	3600	2.00E+09	125.0	0.03	0.07

Table 1. A 1994-2011 MPU summary

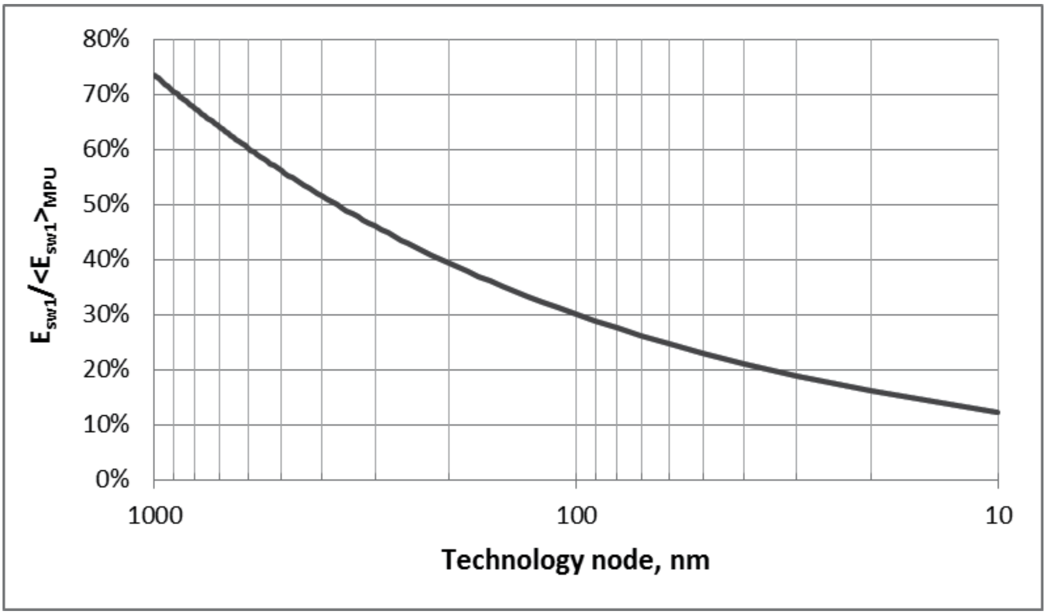


Figure 5. The ratio of the individual transistor dynamic switching energy ('bottom-up') to the average energy per transistor in MPU ('top down').

2.2. ICT Systems

One indicator of the ultimate performance of an information processor, realized as an interconnected system of binary switches, is the binary information throughput (BIT); that is the maximum number of on-chip binary transitions per unit time. BIT is the product of the transistor count N_{tr} with the clock frequency of the microprocessor f:

$$BIT = N_{tr} \cdot f \tag{9}$$

It is instructive to investigate its relation to the overall computational performance of microprocessors, which is often measured in (millions) of instructions per second (IPS) that can be executed against a standard set of benchmarks. As can be seen in Fig. 6, there is a strong correlation between system capability for IPS and the binary throughput, and to a good approximation:

$$IPS = k \times \left(BIT\right)^r \tag{10}$$

For a variety of microprocessor chips (a selection of 39 chips produced in 1971-2011 by 10 different companies, for details, see [11], k~0.1 and r~0.64 with a high degree of accuracy (the determination coefficient $R^2 = 0.98$). This strong correlation suggests a possible fundamental law behind the empirical observation. It is also instrumental for speculations about future developments. According to (10), for a larger computational power, the binary throughput needs to be further increased, which in- turn requires an increase in the number of transistors and/or switching frequency. It is straightforward to show, however, that increasing BIT leads to increased power consumption, according to an equation:

$$P = BIT \times E_{bit} \tag{11}$$

Leading-edge high-performance chips already consume ~100 W of power (Figs. 6 and 7), and this makes their cooling an important issue.

A connection between binary information throughput and power consumption is very visible in memory blocks. Figure 6 shows a linear relation between read power and data rate for several high-speed DRAM systems, consistent with (11).

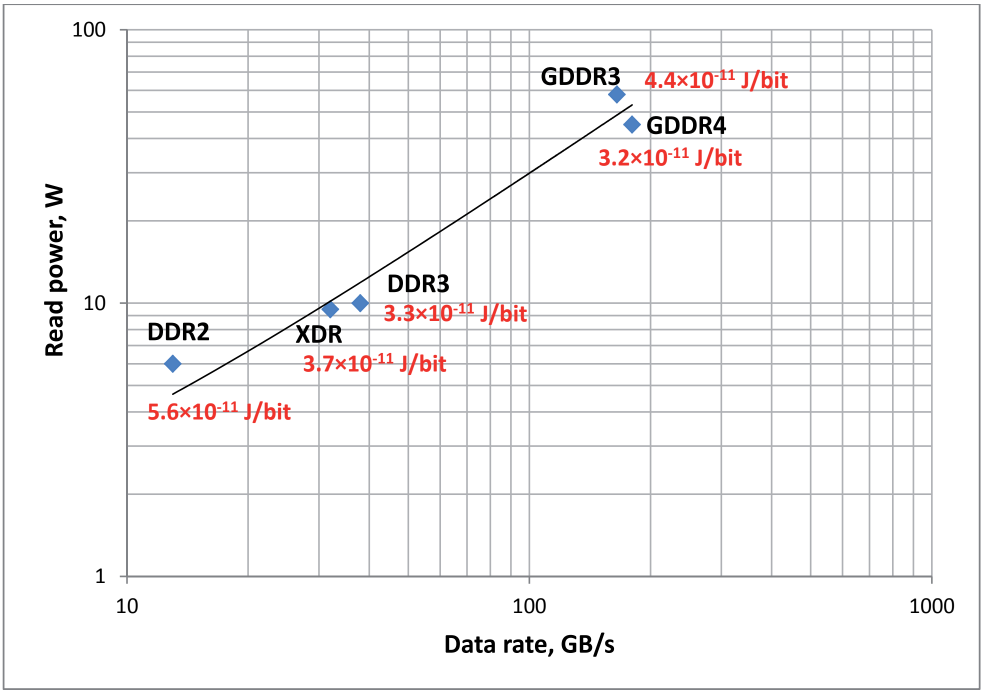


Figure 6. DRAM read power vs. data rate (adapted from [12])

The critical role of wires is also emphasized in memory circuits that are typically organized in arrays connected with long wires shown in an insert in Fig. 2. The wire capacitance (proportional to the wire length) effects system-level energy per bit operation (calculated using (11) and indicated in red in Fig. 6). Note that the DRAM energy per bit is 10,000 times larger than the system-level energy per bit in microprocessors!

When searching for alternative information processing technologies and architectures, the human brain is often proposed as a different model for computation. In [13] an estimate of equivalent binary transitions was made from the analysis of the control function of brain: the equivalent number of binary transitions to support language, deliberate movements, information-controlled functions of the organs, hormone system etc., resulting in an 'effective' binary throughput of the brain ~ 10^{19} bit/s. An estimate of the number of equivalent instructions-per-second (IPS) was made in [14] from the analysis of brain image processing capability resulting in ~10^{14} IPS. It is clear that the brain is not on the microprocessor trajectory in Fig. 7, giving rise to the hope that there may exist alternate technologies and computing architectures offering higher performance (at much lower levels of energy consumption). On the other hand, achieving brain performance with existing ITC would require a massive increase binary throughput of the computing system, and this would also result in high power consumption. For example, the most recent and most impressive demonstration of an artificial intelligence

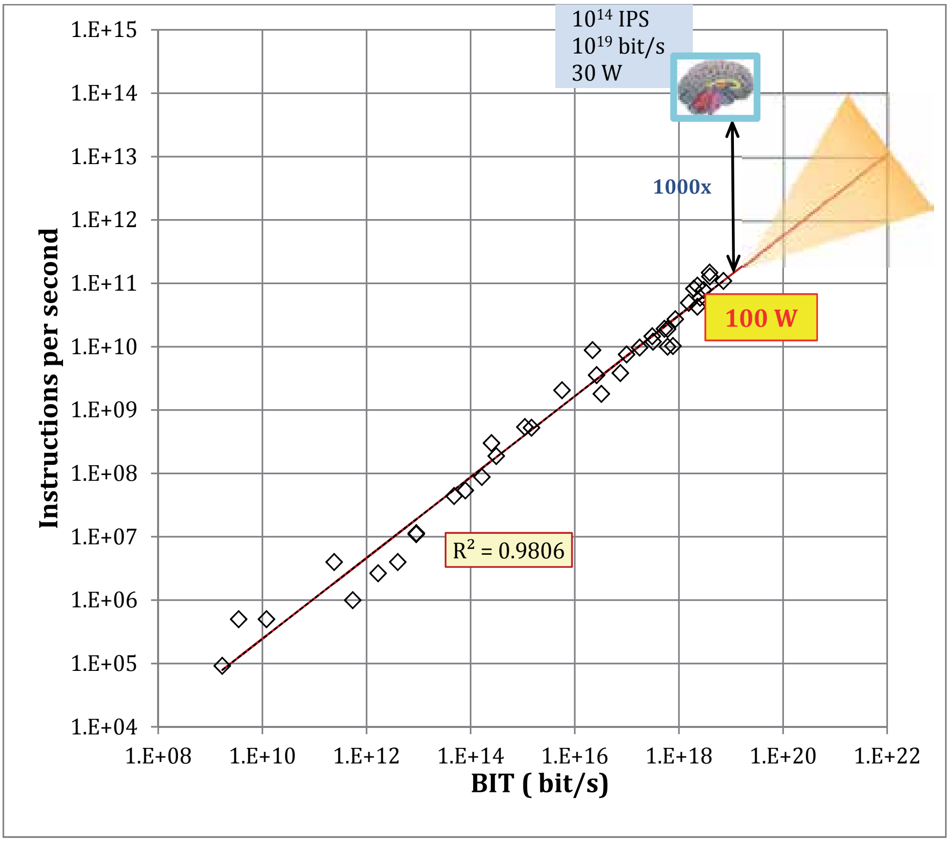


Figure 7. A trend for increasing computational performance through device scaling

computer system, the IBM Watson supercomputer capable of answering questions posed in natural language and the winner of the 2011 *Jeopardy!* quiz show, is built from ~3000 processor cores (POWER7) each consisting of 1.2B transistors and operating at 3.5GHz, thus approximate total binary throughput about 10^{22} bit/s. The machine consumes ~200kW of power [15]. The fact that the brain, a- biological information processor, operates at only ~30W suggests that there may exist alternate technologies and computing architectures offering higher performance (at much lower levels of energy consumption).

3. Fundamental limits in energy dissipation of computing

As we have seen in the previous paragraph, energy dissipation in present computers is an important issue. To reach a better understanding of the basic mechanisms behind energy dissipation in computing devices we propose to approach the energy dissipation issue from a very general and abstract perspective. We start this generalization by considering an ICT device as a black-box machine [16] that performs the activity of processing information by transforming some energy. For the moment we ignore any internal details of the functioning of this black-box. Under this perspective an ICT device can be considered a special thermal machine (see fig. 8).

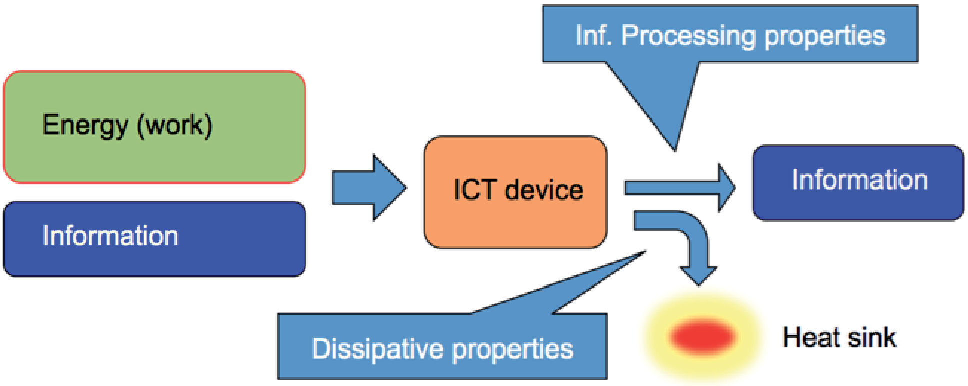
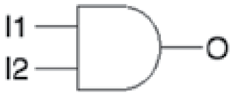


Figure 8. An ICT device is a machine that inputs information and energy (in the form of work), processes both and outputs information and energy (in the form of heat).

A traditional thermal machine is a device that processes energy. More precisely it transforms energy in the form of heat into work for industrial applications. An ICT device is a slightly more complex machine because it processes energy *and* information at the same time. More specifically it inputs a certain amount of information and some energy in the form of work and outputs a reduced amount of information and the same quantity of energy, although in the form of heat (see Fig. 8). In order to define the dissipative processes that take place during its functioning we need to consider both energy and information transformation processes. We have already addressed the energy transformation processes in Chapter 2. Here we focus our attention on the information transformation processes.

3.1. Logic gates

In modern computers the information is processed via networks of *logic gates* that perform all the mathematical operations through assemblies of basic Boolean functions. As an example the NAND gate (Fig. 9) that, due to its universal character, can be widely employed in connected networks to perform a other logic functions. In Fig.10 the combinational networks of NAND gates for performing basic Boolean functions NOT, AND, OR are shown.



I1	I2	O
0	0	0
0	1	0
1	0	0
1	1	1

Figure 9. Symbolic representation of a NAND logic gate and the corresponding truth table. I1 and I2 are the input bits and O is the output bit.

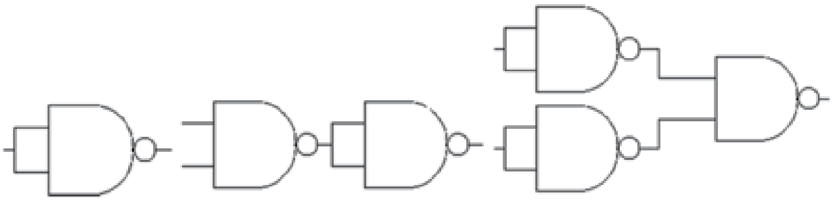


Figure 10. From left to right: NOT gate, AND gate, OR gate, all implemented by connecting together NAND gates.

According to the information preserving role of the logic operations we can distinguish the logic function in *logically reversible* logic gates and *logically irreversible* logic gates. For example the NOT function is implemented by a logically reversible logic gate because the value of the input bit I can be deduced by the value of the output bit O, as it is well evident by inspecting its truth table:

I	O
0	1
1	0

Figure 11. Truth table for the logic gate NOT.

On the other hand the NAND gate (see Fig. 9) is clearly logically irreversible because the knowledge of the output bit O does not allow to one to deduce the value of the input bits I1 and I2 in three cases out of four. According to the definition of *quantity of information* proposed by Shannon[17], an irreversible logic gate decreases the quantity of information at its output. As a simple example consider the NAND gate: there are two bits of information at the input (see truth table in fig. 9) and 1 bit of information at the output. Thus the information balance is negative and the logic gate is irreversible. On the other hand, in the case of the NOT gate there is 1 bit at the input and 1 bit at the output. The information balance is zero and the logic gate is reversible.

This section is entitled "Fundamental limits in energy dissipation of computing" but to date the focus has been on logic gates, i.e. mathematical operations. What has all this to do with energy? To answer this question we have to consider the fact that in a practical computer, the logic gate function is realized by some material system. The bit value is represented by some physical entity (signal) like electric current or voltage, light intensity, magnetic field,…etc. Such signal inputs to the logic gate device and go through a transformation to represent at the output the desired bit values. Modern logic gate devices are made by assembling more elementary units: i.e., transistors. A transistor is an electronic device that performs the role of a switch by letting or not-letting the electric current pass through. Examples of physical implementations of logic gates will be discussed below.

3.2. Landauer limit

In the following section it will be demonstrated that the minimum energy to operate a physical switch can be reduced to zero provided that the amount of information in the switch transformation is not decreased. This condition has been pointed out initially by John von Neumann in a lecture in 1949 [6] and subsequently focussed by R. Landauer [7] and C. H. Bennet[8].

The reasoning is the following: the switch is a macroscopic apparatus composed by many elementary parts (atoms) and thus can be considered a thermodynamic system approximately at equilibrium with the environment. As was shown above, this implies that its transformations are subjected to the laws of thermodynamics. We focus our attention on the single degree of freedom (*dof*) represented by the switch status. This is a dynamical system coupled to the thermal bath represented by all the remaining internal *dof*. A switch event is a change from an initial condition to a final condition. During this change the exchanges of energy and entropy need to be accounted for. If the switch is thermally isolated from the external environment, then there can be no transfer of heat. Suppose for a moment that a switching event can be performed without any work from outside (this point is addressed in the next paragraph), then the only balance that needs to be taken into account is the change in entropy. This change is measured by the change in the macroscopic configuration of the switch. If the change is from state *open* to state *close,* then there is only one initial configuration and one final configuration. There is therefore no net change in the number of configurations and thus no change in entropy according to Boltzmann (see Chapter 2). Let's suppose now that the switch is in an unknown state (it will be shown later that this is the natural condition for a physical switch left alone, after some time), this means that the switch can be in the *open* or in the *closed* state with equal probability in the initial configuration. If now a change is applied to put the switch into a *close* (or *open*, same reasoning) condition, the number of configurations is changed from 2 to 1 and thus there is a change in entropy (Chapter 2) given by:

$$S_f - S_i = k_B(\ln 1 - \log 2) = -k_B \ln 2 \tag{12}$$

where K_B is the Boltzmann constant. The change in entropy is associated with a change in heat via the relation discussed in chapter 2:

$$TdS \geq dQ \tag{13}$$

The equal sign holds when there is no other dissipation associated with the change. Thus based on this reasoning every time that the number of input configurations is smaller than the output configurations, there is a reduction of entropy. Due to the second principle of thermodynamics, this process cannot occur spontaneously and energy expenditure is required. This energy has a lower bound in the amount just given above.

This result can be generalized to ICT devices composed of an arbitrary number of switches. The number of *input configurations* in a network of switches is associated with the *number of input bits* to a system of ICT devices and the number of *output configurations* is related to the number of *output bits*. Thus by computing the quantity of information change during the operation the minimum energy expenditure for the operation can be determined. For the simplest case (sometimes addressed as the *reset operation*) where a switch is set to a given value (*open* or *close*), the minimum energy amounts to:

$$k_B T \ln 2 \tag{14}$$

as anticipated at the beginning of this chapter.

The detailed physics of real switches is discussed in the following sections. The concentration of the following discussion is on energy dissipation and addresses the more fundamental question pertaining to the minimum energy dissipation in *any possible ideal switch*. In this regard a switch is an ideal device that can assume only two states: *open* and *close*.

In the following the focus will be on the physics of a switch with the aim of elucidating the general features associated with energy dissipation mechanisms that take place during the switch operation. In doing this, however, we will ignore those mechanisms that are associated with a specific technology (like the charging or discharging of a capacitor in the electronic realization of a switch) and try to discuss the mechanisms that are common to any possible realization of a physical switch. In order to reach this goal let's start with the definition of switch that we have introduced above: a (bistable) switch is a device that can assume two distinguishable states.

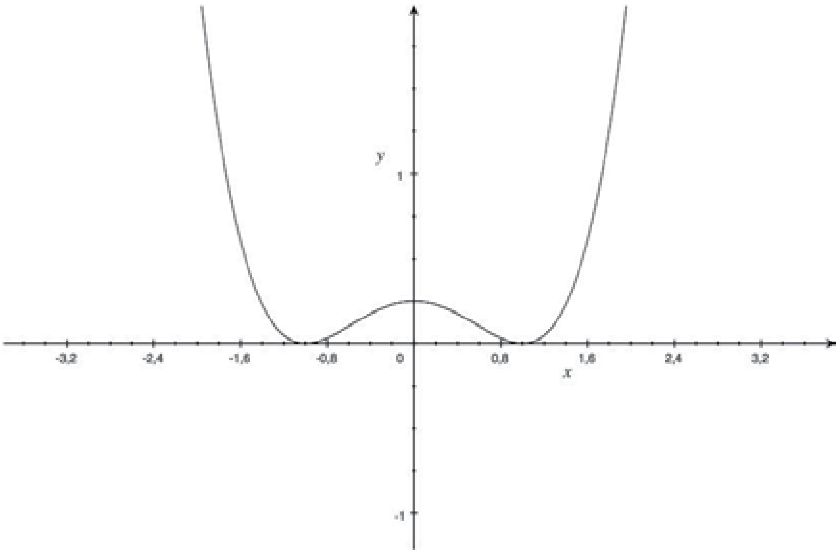


Figure 12. Bistable potential *U (x)*

3.3. The dynamics of a "simple switch"

In order to describe the physics of a switch we need to introduce a dynamical model capable of capturing the main features of a switch, regardless if it is realized with a purely mechanical, electro-mechanical or electronic technology. According to the reasoning originally developed by Landauer[7] we assume that the switch dynamics can be described by a single degree of freedom (*dof*) that is identified with x. Let's suppose that x is a continuous variable (e.g. the position of a cursor or the value of a magnetic field) that can assume two identifiable stable states: e.g. x<0 (logic state "0"), x>0 (logic state "1"). The two states, in order to be dynamically stable, are separated by some energy barrier that should be surpassed in order to perform the switch event. This situation can be mathematically described by a second order differential equation like:

$$m\ddot{x} = -\frac{d}{dx}U(x) - m\gamma\dot{x} + F \tag{15}$$

Where *F* is an external force that can be applied when we want to change state, γ is the frictional force that represent dissipative effects in the switch dynamics and

$$U(x) = -\frac{1}{2}x^2 + \frac{1}{4}x^4 + c \tag{16}$$

is the bistable potential shown in fig. 12. The additive constant, c, is an arbitrary constant that sets the zero level of the potential energy.

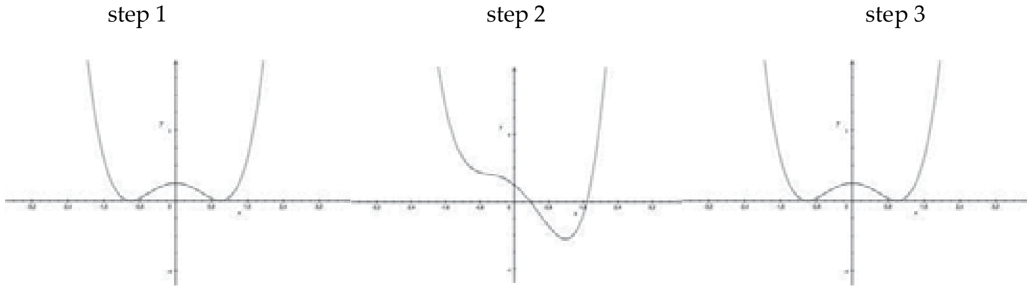


Figure 13. Potential *U (x) + F*. First procedure: From left to right, step 1,2,3. Step 1 and step 3, F=0; step 2, *F=-F*$_0$.

Suppose that at a certain time t_0, the system is x<0 (logic state 0) and *F = 0*. This is equivalent of picturing a material particle of mass *m* and position x, sitting at rest at the minimum of the left well in figure 15 and is an equilibrium condition for the switch.

According to this model if a switch event is to be produced it is necessary to apply an external force *F* capable of bringing the particle from the left well (at rest at the bottom) into the right well (at rest at the bottom). Clearly this can be done in more than one way.

As an example we start discussing what we call the **first procedure**: a three-step procedure based on the application of a large and constant force $F=-F_0$, with $F_0 > 0$.

We start in step 1 (see fig. 13) with the particle on the left well and *F=0*. In step 2 we apply for a certain time $F=-F_0$ in order to change the potential shape into $U(x) - F_0\, x$ (see fig. 13, step 2).

Clearly after some time the particle will move toward the right until it reaches the bottom of right well, where, after few oscillations, it settles due to the presence of the dissipative force. Then, step 3, the force *F* is removed and the system returns to the unperturbed potential of Fig. 12. In this way, a switch event can be produced.

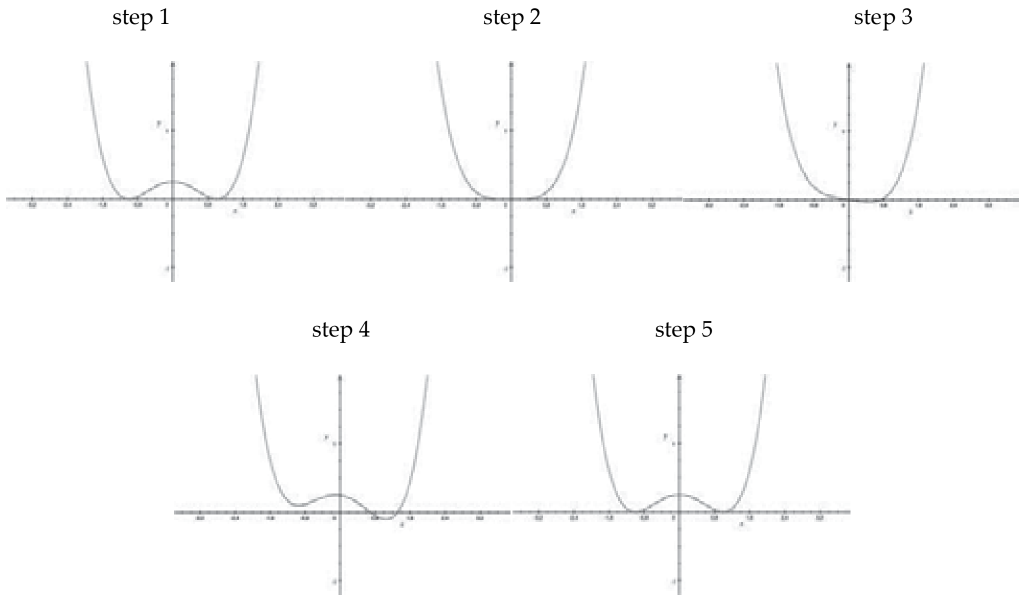
What is the minimum work that the force F must perform to make the device switch from 0 to 1 (or equivalently from 1 to 0). The work is computed as:

$$L = \int_{x_1}^{x_2} F(x)dx \tag{17}$$

where x_1 and x_2 are the starting and ending position of our particle.

In the above example the work is readily computed by considering that the total force acting on the particle is $F_0 + dU/dx$ and has caused a displacement from x_1=-1 to x_2=1. The total work performed is easily computed to be $L_0 = 2\, F_0$. Is this the minimum work? Clearly it is not.

In order to demonstrate that it is possible to switch with a less work, let's consider the following 5-step procedure (**second procedure,** see Fig. 14): in step 1 and step 5 let *F=0;* in step 2 lower the potential barrier by applying a proper force *F=-x*. In step 3 apply an additional small constant force -F_1 that tilts the potential toward the left. Now F=-x-F_1. At this point the material particle slowly moves toward the right. When the particle reaches the far right limit proceed to step 4 and remove the *F=-x* force. Finally in step 5, remove the additional force *F=F_1* and restore the original bistable potential.



step 1

step 2

step 3

step 4

step 5

Figure 14. Potential *U (x) + F*. Second procedure: Step 1 and step 5, F=0; step 2 *F*=-x; step 3 F=-*x*-F_1; step 4 *F*=-F_1;

In order to compute the work performed on the particle observe that in step 1-2 and step 4-5 no work is performed because the applied force does not produce any displacement (or a negligible one). The only work performed happened to be during step 3 where it is readily computed as $L_1 = 2\ F_1$. Now, by the moment that $F_1 << F_0$, as anticipated, we have $L_1 << L_0$. Based on this reasoning it can be concluded that, provided an arbitrarily small constant force is applied during the tilt, the resulting work will be arbitrarily small. Thus it can be concluded that in principle it is possible to perform the switching event by spending zero energy provided two conditions are satisfied: 1) The total work performed on the system by the external force has to be zero. 2) The switch event must proceed with a speed arbitrarily small in order to have arbitrarily small losses due to friction.

3.4. The dynamics of a more realistic "simple switch"

This analysis, although correct, is quite naïve, indeed. The reason is that it has been assumed that the work performed, no matter how small, is completely dissipated by the frictional force. As we have discussed in chapter devoted to energy however, for an isolated system the existence of a dissipative force is the signature of the presence of a large number of degrees of freedom that somehow accommodates the dissipated energy associated with work done by the force. In order to take into account a more realistic representation of the switch dynamics [18] assume that the single-*dof* switch is coupled to a thermal bath that is at temperature T. Although the switch is thermally isolated, exchanges of heat Q between the switch and the thermal bath are possible. Moreover, due to the coupling with the thermal bath a fluctuating force *ξ (t)* appears. At thermal equilibrium the Fluctuation-Dissipation theorem (see Chapter on energy) links *ξ (t)* and the dissipative force. According to this (more physical) description the switch dynamics can be described in terms of a Langevin equation, where the fluctuating force now appears:

$$m\ddot{x} = -\frac{d}{dx}U(x) - m\gamma\dot{x} + \xi(t) + F \tag{18}$$

The fluctuating force *ξ (t)* is represented here by a zero average stochastic process that is defined in statistical terms. The equation of motion has now become a stochastic dynamical equation and its *solution* can be approached in statistical terms. One relevant quantity for describing the system dynamics is represented by the probability density function *P (x,t)*. Specifically *P (x,t)dx* represents the probability for the observable x (the position of the particle) to be at time t within the interval between x and x+dx. Accordingly

$$p_0(t) = \int_{-\infty}^{0} P(x,\ t)dx \ and \ p_1(t) = \int_{0}^{+\infty} P(x,\ t)dx \tag{19}$$

represent the probabilities for the switch to assume the logic states 0 and 1, respectively. As discussed before, it is now possible to address the problem of the work required to perform a switch event in this new thermodynamic framework.

In this case the definition of the switch event itself must be reconsidered. Previously the switch event was defined as the change from an equilibrium position (e.g. at rest at the bottom of the left well) to another equilibrium position (e.g. at rest at the bottom of the right well). In this new thermodynamic framework however the particle is never at rest: due to the presence of the fluctuating force the particle will be randomly oscillating around the potential minima, with occasional random crossings of the potential barrier between the two wells. Since the potential is symmetrical and the fluctuating force has zero average, the two states "0" and "1" have the same probability. This implies that the probability density distribution at equilibrium *P (x,t)=P (x)* is stationary and symmetric, as represented in fig. 15.

Thus if the particle is placed at rest at the bottom of the left well, then after some time t_1 it starts to oscillate around the potential minima and after some longer time t_2 it will jump into the right well and eventually back into the left well and so on. The time t_1 and t_2 are random variables. Their mean values τ_1=<t_1> and τ_2=<t_2> (with $\tau_2 > \tau_1$) can be computed on the bases of the features of the potential *U (x)* and the stochastic force *ξ (t)*. They are usually addressed as the *intra-well* relaxation time and the *inter-well* relaxation time and, roughly speaking they represent respectively the average time the system takes to establish equilibrium within one well (as it would temporarily ignore that the potential is wider than a single well) and the average time it takes to go to global equilibrium. Since τ_2 depends exponentially on the barrier

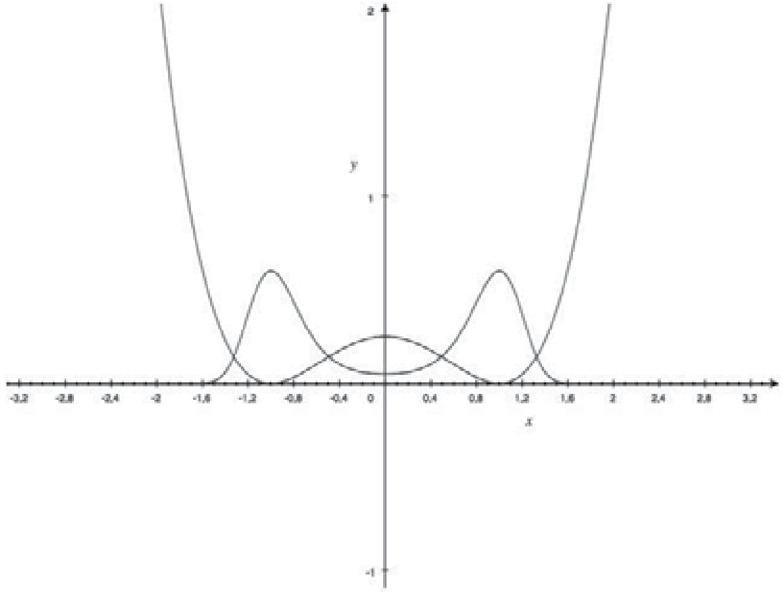


Figure 15. Bistable potential *U (x)* with superimposed the probability distribution P (x,t)=P (x) at equilibrium.

height between the two wells, in practical switches the barrier height is chosen to be large enough to guarantee that $\tau_2 >> \tau_1$.

Based on these considerations define the switch event as the transition from an initial condition toward a final condition, where the initial condition is defined as <x> < 0 and the final condition is defined as <x> > 0. With the initial condition characterized by:

$$p_0(t) = \int_{-\infty}^{0} P(x,\ t)dx \cong 1 \ and \ p_1(t) = \int_{0}^{+\infty} P(x,\ t)dx \cong 0 \tag{20}$$

and the final condition by:

$$p_0(t) = \int_{-\infty}^{0} P(x,\ t)dx \cong 0 \ and \ p_1(t) = \int_{0}^{+\infty} P(x,\ t)dx \cong 1 \tag{21}$$

Clearly the conditions are reversed for a switch event from 1 to 0.

In order to produce the switch event, we proceed as follows: set the initial position at any value x < 0 and wait a time t_a, with $\tau_1 << t_a << \tau_2$, then apply an external force *F* for an elapsed time t_b to produce a change in the <x> value from <x> < 0 to <x> > 0. Then remove the force. In practice it will be necessary to wait a time t_a after the force removal in order to verify that the switch event has occurred, i.e. that <x> > 0. The total time spent has to satisfy the condition $2\ t_a + t_b << \tau_2$.

Now that a switch event has been defined in this new framework, we can return to the question: what is the minimum energy required to produce a switch event?

It is quite easy to see that in order to minimize the energy dissipation the role of the friction has to be negligible. This requires that the switch process must be performed very slowly. As already illustrated in the first procedure described previously (the constant force F_0 procedure) is not optimal. What about the second procedure? We can show that the second procedure, at difference with our previous analysis, although it does allow for a zero work transformation, it does not allow for zero energy expenditure. This is well apparent since in this new thermodynamic framework account must be taken for not only the energy changes due to the external work but also the heat Q passages and thus the role of the entropy S of the system. Based on the discussion in Section 11.3 and more generally in Chapter 2, we have seen that while a switching transformation can be carried-out without spending any energy, a transformation that evolve spontaneously (and thus increases the system entropy), if it is desired to perform a transformation that decreases the system entropy it is necessary to expend a minimum amount of energy $\Delta Q = T\ \Delta S$. In this case (particle in the double well) the system entropy can be computed according to Gibbs as:

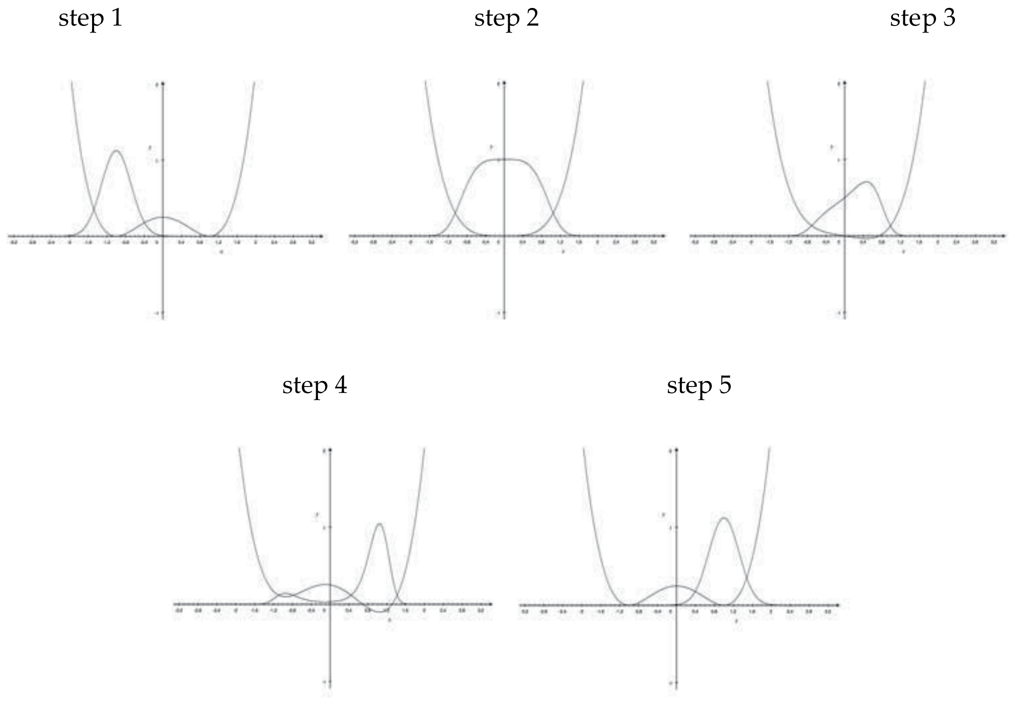
$$S = -\,k_B \sum_i p_i \log p_i \tag{22}$$

Here the sum is limited to the two possible states in our switch and thus i=0,1. Thus if to perform a switch event without spending energy it is necessary to follow a procedure that does not require any entropy decrease during any of the steps. Let's analyse the steps in the second procedure. Note that between step 1 and step 2 the entropy of the system increases. This is due to the fact that the potential is changed by lowering the barrier. At this point the particle dynamics relaxes (in a very short time) to the new configuration and the entropy increases. This is apparent by the change in the probability distribution (see fig. 16) and can be demonstrated quantitatively by simply assuming that in step 1 we have $p_0=1$ and $p_1=0$, this gives $S_1 = -k_B \ln 1 = 0$. In step 2 $p_0=p_1=½$ (see fig. 19) and thus $S_2 = -k_B\ (½ \ln ½ + ½ \ln ½) = k_B \ln 2$. Thus $\Delta S = k_B \ln 2 > 0$. On the other hand, when there is a transition from step 2 to step 5 entropy is reduced from S_2 to $S_5=S_1=0$, thus $\Delta S = -k_B \ln 2 < 0$. According to the thermodynamics theses last steps cannot be performed without providing energy to the system and thus the minimum energy in this case is not zero.

Based on these considerations the conditions required to perform a switching event that expends zero energy can be formulated: 1) The total work performed on the system by the external force has to be zero. 2) The switch event has to proceed with a speed arbitrarily small in order to have arbitrarily small losses due to friction. 3) The system entropy must never decrease during the switch event.

In the following, as an example, a possible procedure (**third procedure**) that satisfies these three conditions is shown. In order to satisfy condition 1), apply a force that keeps the average position of the particle always close to the minimum of the potential well. In this case in fact the force is zero and the work will be zero as well. In order to satisfy condition 2) a change in the applied force should be produced very slowly. Finally in order to satisfy condition 3), i.e., the probability density in state 0 and in state 1 is the same, apply a force that does not change the probability density along the path (constant entropy transformation). This can be done by applying a force that changes the potential as shown in fig. 17. Such a procedure clearly satisfies the three conditions that we enunciated above.

Finally, to conclude this section observe that any physical bistable switch, if left alone for a time that is of the order of τ_2 will eventually evolve into a situation similar to Fig. 15. In this case, when a switch event is required, the operation is completely similar to the reset operation addressed by Landauer in his original works and thus a minimum of k_B T ln 2 is necessarily required to operate the switch.



step 1

step 2

step 3

step 4

step 5

Figure 16. Potential *U (x)* + *F*. Equilibrium P (x) for the different cases. Second procedure: Step 1 and step 5, F=0; step 2 *F*=-x; step 3 F=-*x*-F_1; step 4 *F*=-F_1;

4. Charge based switching devices

Simplest charge based switch (Fig. 18) is a electromechanical device consisting of two metal electrodes, that, depending on the switch's state, are either separated by an air gap (OFF or *open* – non-conducting state) contacts, or touching each other (ON or *closed* – conducting state). The separation between electrodes is changed by applying external mechanical force (e.g.

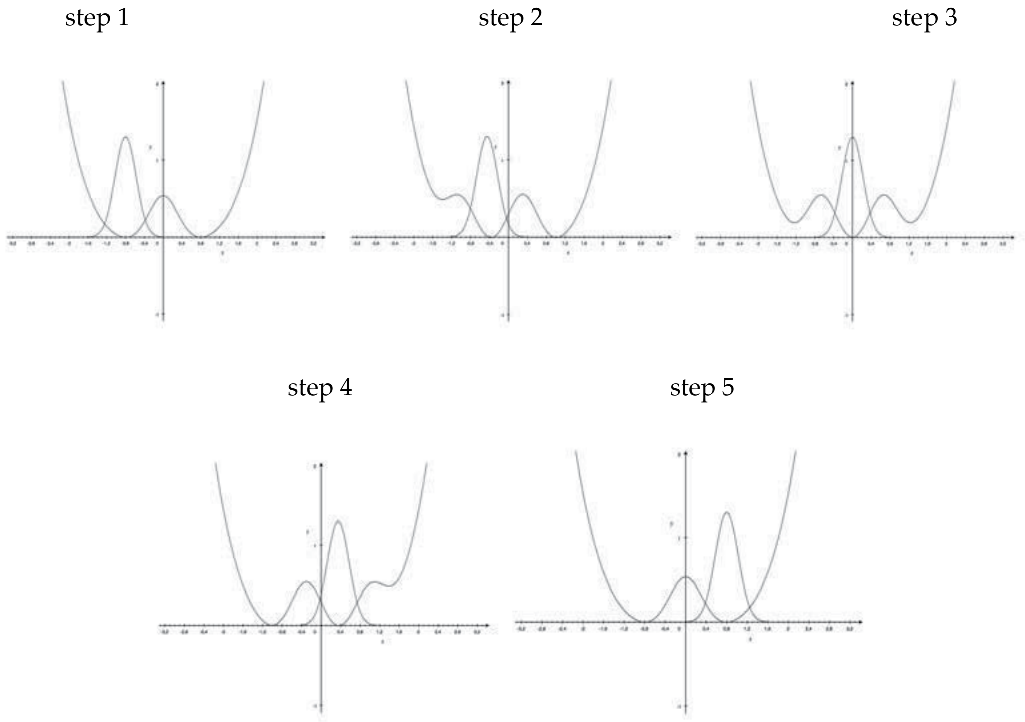


Figure 17. Potential *U (x) + F*. Equilibrium P (x) for the different cases. Third procedure.

manually). Note that in the non-conducting OFF state, an energy barrier is present between the metal electrodes that prevents electron transport between the electrodes (Fig. 18c). A barrier is naturally present at the interface between metal and vacuum (air) and is called work function (WF). A typical work function of stable metals is 4-5 eV. In realistic cases the barrier walls have finite slope and rounded corners due to the image force effect [19, 20]. For smaller gaps, for example when the left-hand electrode is moving towards the right-hand electrode under external field, the shape of the barrier changes, with barrier height reducing and more prominent corner rounding (Fig. 18c). Eventually, the barrier height is reduced to zero (even before the electrodes touch) that manifests the transition to the ON (close) state. The fine transient processes of Fig. 18c are often ignored in a simplified treatment, instead an abrupt transition from a high-barrier to a zero (low)-barrier state is assumed (Fig. 18d).

The bistable switches can be used to implement the three fundamental logical operations, from which all other logic functions, no matter how complex, can be derived. These operations are NOT, AND, and OR. Fig. 19 shows generic schematics for the three basic logic gates. Each logic gate consists of several distinct elements, e.g. switches and resistors. Switches can be implemented by different devices: electromechanical switches and relays, diodes, bipolar or field-effect transistors etc. For example, different Implementations of the NOT gate (inverter) are shown in Figure 20. The generic switch in Figure 20a is implemented by a FET in Figures 20b and c. The resistor in NMOS implementation (Fig. 20b) can also be realized by using a transistor structure.

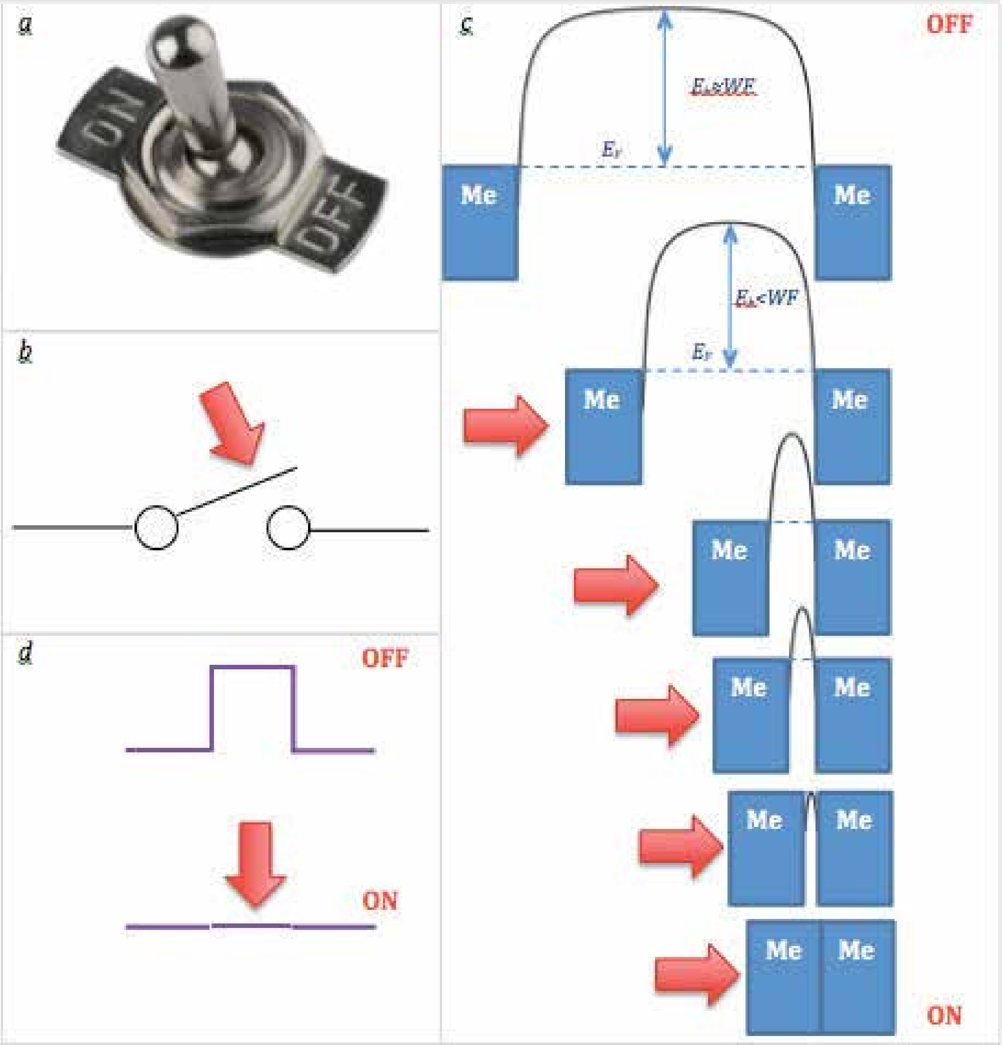


Figure 18. A bistable switch: a) An electromechanical switch example, b) Bistable switch schematics; c) Switch's barrier diagram and its evolution during OFF-ON transition; d) A simplified abrupt barrier transition model

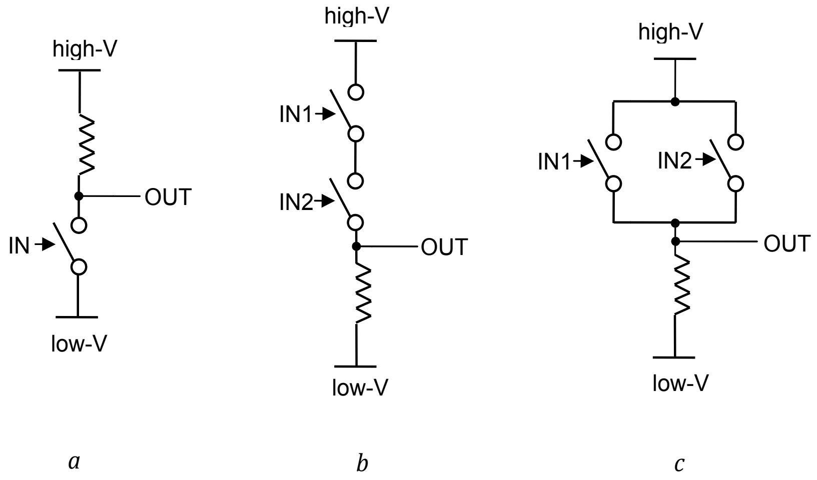


Figure 19. Generic implementations of three fundamental logic operations: (a) NOT; (b) AND, and (c) OR

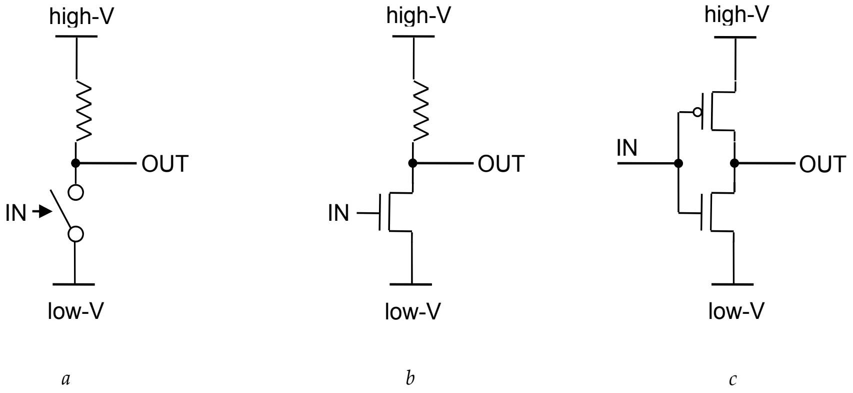


Figure 20. CMOS implementations of three fundamental logic operations: (a) NOT; (b) AND, and (c) OR

Finally in CMOS implementation of Fig. 20c, the resistor is replaced by a "complementary" FET. The role of transistors in ICT is nowadays paramount and the energy dissipation caused by these devices integrated in ICT was outlined in the previous sections.

Two basic electronic devices for information ICT will be considered here: the binary logic switch and the binary memory element. As was shown in previous sections, the *controllable barrier model* is an useful abstraction for these devices that allows for a simple and intuitive analysis of the physics-based operational limits. At least one energy barrier is always present in ICT devices, and it is fundamentally linked to the nature of information, which is a measure of *physically distinguishable states* 21. If specified locations of an information-bearing particle (e.g. electron) are used to define distinguishable states, a barrier is needed to prevent spontaneous transitions of an information–bearing particle from its 'prescribed' location (Fig. 21). The barrier must also be controllable, i.e. there must be a certain *gating* mechanism to reduce (ideally to zero) its height (or width) to allow wanted transition between informational states. Thus the three generic requirements for the implementation of a particle-based binary switch are i) the ability to detect the presence/absence of the particle in e.g., the location '0' or '1', ii) the ability to preserve on demand the particle in the location '0' or '1', and iii) the ability to move on demand the particle from '0' to '1' and from '1' to '0'.

4.1. Electronic switches

In the above, time-dependent controlled barrier transitions based on gradual adjustments of barrier shape and height were considered. In most practical cases, the treatment can be simplified assuming the barrier transitions are necessarily fast and abrupt as shown in Fig 21. Using this abrupt transition model, we offer below a quick snapshot of the current state and limitations of the electronic computing technologies due to thermal noise and quantum fluctuations. An elementary switching operation of a binary switch consists of three distinct phases shown in Fig. 21. For example, consider the switch in Fig. 21a switching from "0" to "1". The three steps are: 1) An external gating stimulus (e.g. voltage in case of charge-based devices) is applied to the barrier to reduce it from E_b to 0, 2) the particle moves from'0' to '1' location (for this transition, an additional kinetic energy E_k must be supplied to the particle), and 3) the barrier height is restored back from 0 to E_b to preserve the final '1' state. All three modes have characteristic times determined by physics and can be described by the coordinate and velocity of the information carrier/material particle, and by corresponding energies. The work required to suppress or restore the barrier is equal or larger than $E_{b.}$ It is important to note that in electronic devices, for technological reasons, this work is considered lost energy, a condition that was not present in our previous ideal model in 11.3, when the lowering or raising the barrier did not required per sè a finite amount of energy. Specifically in electric charge based devices, changes in the barrier height require changes in charge density, and as a result this always requires charging or discharging of a certain capacitor. As we have discussed above, this require a certain amount of energy dissipated. Thus, in the first order the barrier height determines the energetics of the ICT devices and it is desirable to keep E_b as low as possible for low-energy operations. How small can the energy height be? The energy barrier is needed to preserve a binary state in the presence of fluctuations, both classical (thermal noise) and quantum effect (tunneling) are present. In the following we briefly discuss these two important aspects.

Thermal noise.

The thermal noise is directly related to the fundamental result of thermodynamics, which states that each material particle at equilibrium with the environment possesses kinetic energy of ½ $k_B T$ per degree of freedom due to thermal interactions, where k_B is the Boltzmann's constant and T is absolute temperature. The permanent supply of thermal energy to the system occurs via mechanical vibrations of atoms (phonons) and via the thermal electromagnetic field of photons (background radiation). Thus the barrier height, E_b, must be large enough to prevent spontaneous transitions (errors) [22] that occur when the particle spontaneously acquires

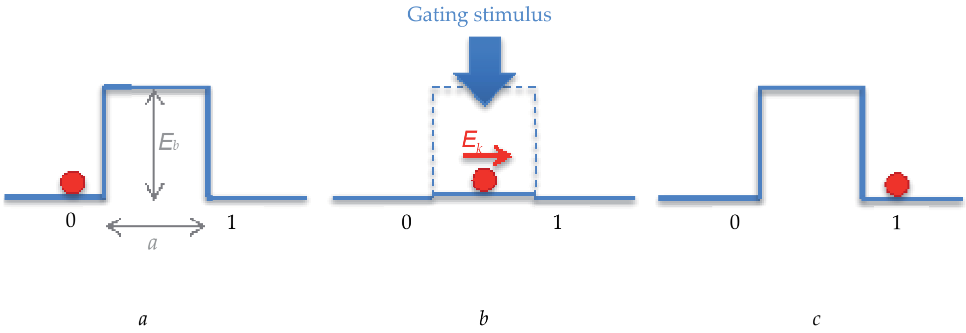


Figure 21. Three phases of abrupt binary switching

thermal energy large enough to jump over the barrier. This can easily happen if the kinetic energy of the particle E is larger than the barrier height E_b. This can be easily seen, by estimating the probability for over-barrier transition from the Boltzmann distribution:

$$f(E) = A\exp\left(-\frac{E_b}{k_B T}\right) \tag{23}$$

The probability of over-barrier transition is equivalent to the probability that the particle gains energy $E>E_b$ which probability is obtained by integration of (23):

$$p = \int_{E_b}^{\infty} f(E)\, dE = A\int_{E_b}^{\infty} \exp\left(-\frac{E}{k_B T}\right) dE = Ak_B T\exp\left(-\frac{E_b}{k_B T}\right) \tag{24}$$

The coefficient A can be found from the normalization condition for (23):

$$\begin{gathered} 1 = \int_{0}^{\infty} f(E)\, dE = A\int_{0}^{\infty} \exp\left(-\frac{E}{k_B T}\right) dE = Ak_B T = 1 \\ A = \frac{1}{k_B T} \end{gathered} \tag{25}$$

Substituting (25) into (24) obtain that the probability of over-barrier transition is

$$p_{o-b} = \exp\left(-\frac{E_b}{k_B T}\right) \tag{26}$$

The minimum barrier height can be found from the distinguishability condition, which requires that the probability of errors $p < 0.5$, in which case the switch is being operated at the threshold of distinguishability. Solving (24) for p = 0.5, obtain the Boltzmann's limit for the minimum barrier height, E_{bmin} as

$$E_{b\min} = k_B T \ln 2 \approx 0.7 k_B T \sim k_B T \tag{27}$$

Of course error probabilities much less than 0.5 are required in practice, and therefore the barrier height E_b must be larger. For example in modern DRAM the probability of one erroneous bit is ~10^{-9} in a month [10].

The barrier model along with (23) can be further applied to derive the classic formula for the thermal (Nyquist-Johnson) noise, which plays a fundamental role in analog devices:

$$\left\langle V_n^2 \right\rangle = 4 k_B T R \Delta f \tag{28}$$

($\left\langle V_n^2 \right\rangle$is the variance of the noise voltage across a resistor due to thermal agitations, R is the resistance and Δf is operational bandwidth. A derivation of (28) using the barrier model is considered in [23]).

Quantum effects

Another class of errors that impose limits on device scaling are quantum errors, which occur due to quantum mechanical effects. These effects play a measurable role in a system whose energy (*E*), momentum (*p*), space (*l*) and time (*t*) are such that the characteristic physical parameter, the action, $S \sim E\,t \sim p\,l$, is comparable to the *quantum of action* h=6.63×10^{-34} J s (Planck's constant). The corresponding relations are known as *Heisenberg Uncertainty Principle*:

$$\Delta x \cdot \Delta p \sim \frac{h}{2} \tag{29}$$

From (29), the minimum size of a scaled computational element or switch (Fig. 21) is

$$L_{\min} > \Delta x \sim \frac{h}{2\Delta p} = \frac{h}{2\sqrt{2mE_b}} \tag{30}$$

where m is the mass of the information-bearing particle, for example that of the electron.

The Heisenberg relation (29) and its derivative (30) can be used for an elementary derivation of an analytical form of tunnelling probability (known as Wentzel-Kramers-Brillouin (WKB) approximation):

$$p_{WKB} \sim \exp\left(-\frac{2\sqrt{2m}}{\mathrm{h}} \cdot a \cdot \sqrt{E_b} \right) \tag{31}$$

Note that (30) and (31) emphasizes the parameters controlling the tunnelling process. They are the barrier height E_b and barrier width a as well as the mass m of the information-bearing particle. If separation between two wells is less than L_{min} (30) the barrier structure of Fig. 21a would allow significant tunnelling, which will destroy the binary information. Also, parasitic leakage current will considerably increase the total power consumption. For a numerical example using E_b=0.1eV and the effective mass of electron in semiconductor m^*=0.19m_e (the transverse electron effective mass in Si) obtain from (30)

$$L_{\min} \sim \frac{6.63 \cdot 10^{-34}}{2\sqrt{2 \cdot 0.19 \cdot 9.11 \cdot 10^{-31} \cdot 0.5 \cdot 1.6 \cdot 10^{-19}}} \approx 4.5nm,$$

which is an approximate minimum channel length of the Si *logic* FET[21]. This assessment is consistent with ITRS, which projects the minimal physical gate length in logic FET to be ~5 nm [10]. At this scale, leakage due to quantum mechanical tunnelling will be very significant and may limit usage of these ‘ultimate’ devices in many practical applications.

4.2. Memory elements

Next consider ultimate dimensional scaling of the memory elements. To estimate the needed barrier properties for memory, one needs to understand the limits on electrical conductance, which can be done using another form of the Heisenberg relations [24]:

$$\Delta E \Delta t = \frac{h}{2} \tag{32}$$

Let’s consider an elementary act of electrical conductance for an electron passing from reservoir **A** with energy E_A to reservoir **B** with energy E_B (Fig. 22).

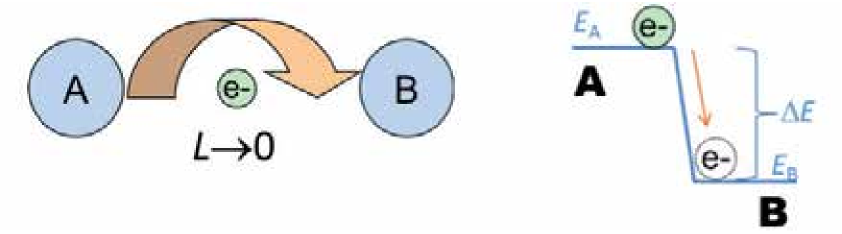


Figure 22. Illustration to the derivation of quantum conductance

The corresponding voltage (potential difference) between **A** and **B,** V_{AB} and the current, $I_{AB,}$ flowing from **A** to **B** are:

$$|V_{AB}| = \frac{E_A - E_B}{e} = \frac{|\Delta E|}{e} \tag{33}$$

$$I_{AB} = \frac{e}{\Delta t} \tag{34}$$

The minimum passage time Δt from (32) is:

$$\Delta t = \frac{h}{2\Delta E} = \frac{h}{2eV} \tag{35}$$

Putting (35) into (34), and taking into account Ohm's law, i.e. I=V/R, obtain:

$$I_{AB} = \frac{2e^2}{h} \cdot V = \frac{V}{R_0} \tag{36}$$

where

$$R_0 = \frac{h}{2e^2} = 12.9k\Omega \tag{37}$$

is quantum resistance. A related parameter is quantum conductance:

$$G_0 = \frac{1}{R_0} = \frac{2e^2}{h} \tag{38}$$

The quantum resistance/conductance sets the limit on electrical conductance in a one-electron channel *in the absence of barriers*.

$$I_0 = \frac{V}{R_0} = \frac{V}{12.9k\Omega} \tag{39}$$

If a barrier is present in the electron transport system, the conductance will be decreased due to the barrier transmission probability $p_T < 1$. The electrical conductance in the presence of barrier is obtained by multiplying the barrier-less quantum conductance (38) by the barrier transmission probability:

$$G = \frac{1}{R} = G_0 \cdot p_T \tag{40}$$

Eq. (40) is a form of the *Landauer formula* [25] for a one-electron conductive channel.

Let now consider as an example the insulator-conductor-insulator memory element shown in Fig. 23, which is representative of floating gate cell used in flash memory. In memory cells that store electron charge, two distinguishable states 0 and 1 are created by the presence (e.g. state 0) or absence (e. g. state 1) of electrons in the charge storage node. In order to prevent losses of the stored charge, the storage node is defined by energy barriers of sufficient height E_b to retain charge (Fig. 23). Assume only one electron is stored. The store time (or corresponding characteristic escape time) is:

$$t_s = \frac{e}{I_{leak}} \tag{41}$$

The two mechanisms of the charge loss are over-barrier leakage and through-barrier tunnel leakage. In both cases the leakage current from the storage node can be calculated from the Landauer formula (40):

$$I_{leak} = G_0 \cdot V \cdot p_T \tag{42}$$

In the following, the thermal voltage $V = \frac{k_B T}{2e}$ will be used as a lower bond.

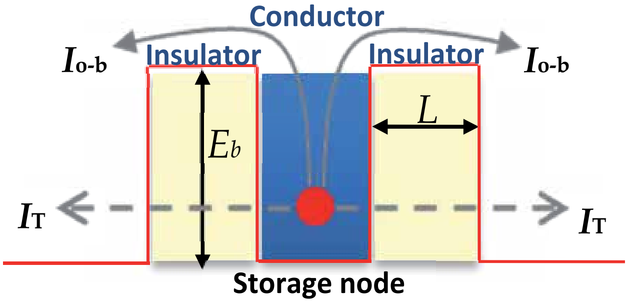


Figure 23. An insulator-conductor-insulator memory element, representative of flash memory.

The probability of thermal over-barrier transitions is the Boltzmann probability - From (42) and (26) the one-electron over-barrier current $I_{o\text{-}b}$ is:

$$I_{o-b} = 2\frac{e}{h} \cdot k_B T \cdot \exp\left(-\frac{E_b}{k_B T}\right) \tag{43}$$

The factor of 2 in (43) appears because escape is possible over either of two barriers that confine an electron as shown in Fig.23.

The electron escape time (the retention time) due to over-barrier transport is:

$$t_{o-b} = \frac{h}{2k_B T} \exp\left(\frac{E_b}{k_B T} \right) \tag{44}$$

If over-barrier leakage is the only mechanism of charge loss (when the barrier width a is sufficient to suppress tunneling), the escape time is equal to the one-electron retention time, $t_{o\text{-}b}$=t_s.

For a specified t_r, the required minimum barrier height is:

$$E_{b\min} = k_B T \ln\left(\frac{2k_B T}{h} t_s \right) \tag{45}$$

In the case of the 'minimum nonvolatile memory' requirement, i.e. t_r>10 years, (45) gives E_{bmin} ≥ 1.29 eV (~50k_BT) at T=300 K.

A second source of charge loss is electron tunneling. The corresponding tunneling current I_T is:

$$I_T = 2\frac{e}{h} \cdot k_B T \cdot \exp\left(-\frac{2\sqrt{2m}}{\mathrm{h}} \cdot a \cdot \sqrt{E_b} \right) \tag{46}$$

The electron escape time due to tunneling is:

$$t_T = \frac{h}{2k_B T} \exp\left(\frac{2\sqrt{2m}}{\mathrm{h}} \cdot a \cdot \sqrt{E_b} \right) \tag{47}$$

The total retention time due to both mechanisms can be estimated as:

$$t_r = \frac{e}{I_{o-b} + I_T} \tag{48}$$

Suppose that the barrier height is large enough to suppress over-barrier escape, i.e. E_b>>E_{bmin}, where E_{bmin} is given by (45). In this case, the store time will be determined by the tunneling time, t_T: t_s≈ t_T. The minimum barrier width for a specified store time, can be estimated from (47), e.g. for t_s=10 years:

$$a_{\min} = \frac{\mathrm{h}}{2\sqrt{2mE_b}} \ln \frac{k_B T}{h} t_s \tag{49}$$

As a numerical estimate for t_s>10 years, E_{bmin}≥ 1.29 eV, m=m_e and T=300 K, (48) gives $a_{\min}$ ~ 5 nm.

As follows from the above, in order to obtain a nonvolatile electronic memory cell, sufficiently high barriers must be created to retain the charge for a long period of time. If different practical factors are taken into account (such higher temperature e.g. T=400 K, lower effective electron mass in solids, e.g. m^*= 0.5 $m_{0,}$ many-electron distribution in solids etc,. for >10 y retention, the minimal barrier height E_b is ~2 eV (~77 k_BT), and thickness a>5 nm [26]. The corresponding practical minimum size of the floating gate cell is ~10 nm [26]. Large barriers also result in high voltages required for memory operation: ~5 V for READ and ~15 V for WRITE [26].

4.3. Energy per bit operation

As it was mentioned earlier, in charge-based devices, changes in the barrier height require changes in charge density, and as a result this always requires charging or discharging of a certain capacitor associated with the device (e.g. a gate capacitor C_g in the case of FET, interconnect line capacitance C_{line} etc.). It was shown in section 11.2 that when a capacitor C is charged from a constant voltage power supply, the energy of $CV^2/2$ is dissipated, and operation of binary devices in this regime is sometimes referred as *irreversible* switching. The total energy per bit operation depends on the device barrier height E_b and the number of electrons N_e involved in the switching process. The minimum energy needed to suppress the barrier (e.g. by charging the gate capacitor) is equal to the barrier height E_b. Restoration of the barrier (e.g. by discharging gate capacitance) also requires a minimum energy expenditure of E_b. Thus the minimum energy required for a full switching cycle is at least $2E_b$. Additional kinetic energy E_k (typically ~E_b) also needs to be supplied to electrons to enable the transition, If N_e is the number of electrons involved in the switching transition between two wells, the total minimum switching energy is

$$E_{bit\min} = 2E_b + N_e E_k = (N_e + 2)k_B T. \tag{50}$$

If N_e=1,

$$E_{bit\min} = 3k_B T \approx 10^{-20} J, \tag{51}$$

and this is a lower boundary for a logic operation. For a nonvolatile memory device, that requires a minimum barrier height E_b~$50k_BT$ the minimal energy to store one electron is E_{bit} ~ 150 k_BT.

The above analysis considered individual logic and memory devices operating in a single electron limit. In practice, a larger number of electrons, N_{el}, is needed to support communication between different devices in the system. For logic operations, one 'upstream' binary switch controls/communicates with several 'downstream' binary switches. The number of the downstream devices that are driven by a given upstream device is called 'fan-out' (FO). A typical fan-out in the baseline microprocessors is four (FO4). For communication, the devices are interconnected with metal wires, and in a 2D layout at least one electron needs to be sent to each of the 'downstream' gates, thus, at least four electrons needs to be provided by the 'upstream' devices, and according to (50) $E_{bit} > 6\ k_BT$. In practice, the number of electrons is much larger to ensure communication reliability. Next, at least a few long interconnects are needed to ensure communication between the information processing system and the outside world (e.g. I/O). The energy costs associated with long interconnects can be estimated as energy needed to charge/discharge a metal line of length L:

$$E \sim C_{line}\ V^2 \tag{52}$$

Using line capacitance of C_{line}~$\varepsilon_0 L$ and nearly minimal distinguishable voltage: V~k_BT/e obtain as a lower boundary for the communication energy per unit length:

$$E > \varepsilon_0 \left(\frac{k_B T}{e} \right)^2 \sim 10^{-14}\ \frac{J}{bit \cdot m} \tag{53}$$

For an example of a long wire along a 1 cm chip the limiting communication energy is 10^{-16} J/bit, i.e. 10,000x times more than the minimal energy required for computation!

The long wire considerations are critical for memory that typically is organized in regular X-Y arrays of memory cells. In many instances the properties of interconnecting array wires determine the operational characteristics of the memory system. A given cell in an array is selected (e.g. for read operation) by applying appropriate signals to both interconnect lines, thus charging them. The relatively large operating voltage of flash results in rather large line charging energy. For a memory cell pitch of 10nm and a 128×128 array the line capacitance is ~10^{-14} F [27].For write operation with V_{write}~15V, the write energy is ~10^{-12} J/line. (In practical flash memory devices the read energy is of the order of 10^{-13}-10^{-11} J/bit read and 10^{-9}-10^{-10} J/bit write [28, 29]).To summarize, electrons flowing in metal wires constitutes the main component in energy consumption in the electron based devices. It can also be argued that fixed wiring is among the main factors limiting the efficiency of computational systems. Energy consumption (per bit) in different components of ICT as discussed in this chapter is summarized in Table 2.

5. Open issues and conclusions

Even with steady decreases in the energy required to switch a bit as shown in Fig. 4, it appears that the 'effective' energy required to switch a bit is decreasing at a slower pace. The other essential components of an information processing system are assuming a relatively more significant role in system energy consumption. For example, increases in energy utilization by I/O systems, increases in memory access energy costs due to the array structure of the memory architectures, increases in power consumption by the internal chip wires, device leakage in the OFF state, etc., are consuming a greater share of information processing system energy.

	Fundamental limit	Baseline technology
	Logic	
Barrier height	$k_BT\ln2=18\text{meV}= 3\times10^{-21}\text{J}$	$10^{-19}\text{J} \sim 0.5\text{-}1\text{eV}\sim 24k_BT$
Logic device	$\sim3k_BT \approx 80\text{meV} \approx 10^{-20}\text{ J}$	$3\times10^{-17}\text{J}^* \approx 188\text{eV} \approx 7{,}250k_BT$
Logic circuit	$>6k_BT \approx 160\text{meV} \approx 2\times10^{-20}\text{J}$	$10^{-16}\text{J}^* \approx 625\text{eV} \approx 24{,}150k_BT$
	Nonvolatile Memory	
Barrier height	$\sim50k_B\text{T}\sim1.3\text{eV}\sim2\times10^{-19}\text{J}$	$\sim77\text{k}_\text{B}\text{T} \sim 3\times10^{-19}\text{J}$
Memory device	$\sim150k_BT\sim4\text{eV}\sim6\times10^{-19}\text{J}$	$\sim230\text{ k}_\text{B}\text{T} \sim 10^{-18}\text{ J}$
Memory array	$2\times10^5\ k_BT\sim10^{-15}\text{J}$	$10^{-11}\text{-}10^{-13}\text{J}$
I/O	10^{-16} J	10^{-11} J

*see Table 1

Table 2. Energy consumption (per bit) in ICT: A summary

Leading edge devices today utilize slightly over three orders of magnitude more energy than the k_BT ln2 thermodynamic limit. If current trends continue, it is likely that further reduction in the energy per bit of a device will not be accompanied by corresponding decreases in effective energy-per-bit when viewed at the system level. Moreover, a second issue associated with continuing scaling of device features and supply voltages is that thermal and tunneling noise will require increased use of error correction mechanisms.

It has been observed that it might be possible to operate a switch at energy close to k_BT ln2 for irreversible switching procedures and even lower for entropy preserving switching procedures. This possibility was examined and shown to be theoretically possible; however, the side attributes associated with achieving such a functional device may not be acceptable in practice. For example, there is a need for very slow operation of the device that may be untenable and the energy recovery mechanisms associated with energy storage and retrieval are difficult to implement without incurring energy loss. However, even if one assumes that this can be achieved without any overhead penalties, the communication and fan-out cost of interconnects and I/Os may make this achievement almost invisible. It is also unlikely that energetics of memory devices can be significantly changed. These limitations are even more apparent in charge based devices where *the main source of energy consumption is due to electrical charging large capacitances in metal wires*.

Having said all that, extremely low energy computation may be achievable in systems nased on fifferent technologies. One example is represented by living systems, where it has now been established that individual cells, the smallest units of living matter, possess amazing computational capabilities, and are indeed the smallest known information processors [30, 31]. As argued in a number of studies, individual living cells, e.g. bacteria, have the attributes of a Turing Machine, capable of a general-purpose computation [30, 32, 33], and von Neumann's Universal Constructor, i.e. *computer making computers*[32] (DNA molecule acts as nonvolatile memory of the cell computer, while many proteins in cell's cytoplasm have as their primary function the transfer and processing of information, and are therefore can be regarded as logical elements of the biological cell processor [34, 35, 36, 37]). The Universal Constructor model is a useful concept for the estimation of the information content of a living cell, for example for the *E.coli* bacterium the estimated information content is ~10^{11}-10^{12} bit [11, 23] (interestingly, experimental entropy reduction measurements of the informational content of bacterial cells using microcalorimetric techniques yielded very similar results [38]). Assuming a conservative edge of cell's information content, which is 10^{11} bit and ~ 3000 s for reproduction time of a bacterial cell obtain ~10^7 equivalent bits that must be processed per second (equivalent binary throughput). The power consumption of *E.coli* is about 1.4×10^{-13} W so that from (11) the energy per equivalent binary operation in the cell can be calculated to be ~10^{-20} J or <10 k_BT. Note, that this is the total energy per bit, taking into account logic, memory, I/O (e.g. sensing, ribosomial synthesis etc.).

	Biological Cell Processor	Baseline ICT
Memory	10^7 bit	10^7
Logic	10^6 bit	10^6
Energy per bit	~10^{-20} J(average system level)	10^{-13} (Memory) 10^{-16} (Logic)
Binary throughput	10^7 bit/s	10^7 bit/s
Task time	3000 s	3000 s
Total energy per task	10^{-10} J	10^{-6} J

Table 3. Comparison of the two technologies

The estimated energy utilization per switching event is quite impressive. It can be compared to an equivalent electronic system consisting of the same number of logic and memory elements implemented in baseline technology (Table 2). A comparison of the two technologies reveals that the biological cell processor operates with the four orders of magnitude lower energy than the baseline electronic processor (Table 3).

What makes biological cell a superior information processor relative to the performance of ultimately scaled semiconductor technology? It appears that several simple physics based arguments can be made:

1. Heavier mass of information carrier allows for denser logic and memory. As was argued in section 11.4, a heavier mass for the information carrier allows for smaller separation between distinguishable states and therefore more devices/states per unit volume or area. According to (30), a heavier mass results in smaller device size. For example, DNA memory uses molecular fragments (nucleotides) as information carriers, each consisting of more than 10 atoms. The molecular information carriers are densely packed in a linear array with distance between nucleotides of only 0.34 nm. By comparison, the minimum size of an electron memory cell is ~10 nm. This dimensional difference gives insight into the 1000x difference in volumetric memory density between electronic and DNA memory, i.e. 10^{16} bit/cm^3 of electronic memory vs. 10^{19} bit/cm^3 for DNA memory. Also, protein logic devices consist of arrangements of many atoms, resulting in total device size of ~5 nm or less, which can explain about 10x higher protein logic density compared to ultimately scaled transistors. In fact, a vision of cellular enzyme proteins as conformon-based "soft-

state nanotransistors" has recently been recently introduced in a book by Ji and contrasted with electron-based solid-state transistors [39].

2. Utilization of ambient thermal energy allows for energy minimization in logic circuits. For semiconductor systems thermal energy (~k_BT) must be managed as it may destroy the state or divert the information carrier from its intended trajectory; for example in communication between several logic elements. In order to overcome the deleterious effects of thermal energy each logic element must contain a barrier E_b > k_BT. Moreover, in communication with other elements, N carriers must be sent to the recipient elements, each of which must have kinetic energy Ek > k_BT. As result the total energy per bit operation, as it was derived in section 11.5, becomes (50):

$$E_{bit_{\min}} = 2E_b + N_e E_k = (N_e + 2)k_B T$$

and it can be significantly large, usually >1000 k_BT.

In contrast, biomolecular computing systems utilize thermal energy to effect data exchange/transmission between e.g. logic-to-logic or memory-to-logic elements. All computational molecules move within the cell's volume by thermally excited quasi-random walk with almost no extra energy required, thus E_k~k_BT and the second term is minimized. Biological systems actually use thermal energy in the transmission of information and in the realization of work-related tasks40. Examples of beneficial use of non-equilibrium fluctuations are present also at micro and nano scale level: see e.g. the paradigmatic phenomenon of Stochastic Resonance41.

3. Flexible/on-demand 3D connections/routing allow for minimization of the communication carriers. Referring to (50), in silicon systems most energy is consumed by interconnect. This is due to the need to pump a large number, N, of carriers (electrons) into the interconnecting wire for reliable communications. As was argued in section 11.5, for reliable communication, N must dramatically increase for longer path lengths and more receiving devices (fan out). In electrical circuits the connection paths are pre-determined in 2-D networks, and in many instances, the electron travels a long distance. A problem of electrical interconnects is the statistical behavior of discrete charges, in other words electrons are free to move along the line. Therefore a large number of electrons is needed for reliable branched communication to reduce thermal and shot noise. Electrons flowing in 2-D networks of metal wires constitute the main component in energy consumption in the electron based systems. In contrast, 'devices' in cells (e.g. proteins or RNA) are usually free to travel in all three dimensions within the cell and they don't follow a fixed path. Due to the shape-specific molecular recognition (e.g., lock-and-key interactions) a 'deterministic' or 'point-to-point' communication of information packages within the processor is obtained, with smaller number of carriers, resulting on lower energy.

4. Array-free organization of DNA memory enables the minimization of the energy for memory access. A core system-level challenge resulting in the excessive energy consumption in the silicon microcomputer is that memory access to support computations takes too much energy. Organizing solid-state memory in cross-bar arrays, while an elegant

solution at larger scale, contributes to excessive energy dissipation due to line charging during memory access as given by:

$$E \sim C_{line} V^{2}$$

The long wires needed to connect memory elements in an array result in a large line capacitance C_{line}, which together with a large access voltage required for nonvolatile electron-based memory, yielding 10^{-13}-10^{-11} J per randomly accessed bit. In contrast, the DNA memory in the cell uses array-less organization that can be viewed as similar to access to tape or hard disc drive. Multiple read heads (formed by RNA polymerase protein) are used for independent simultaneous access to different parts of the DNA memory, thus this is a highly parallel process.

5. Hybrid digital & analog information processing. As it is argued in [36] the cell processor is a hybrid state machine operating in both digital and analog modes. For example, DNA memory is a digital unit while the sensory information the cell receives from its environment is mostly analog. The protein-based computing often represents and processes information in analog form, with state variables encoded in concentrations of protein molecules.

From the above discussion, it would appear that the performance and energy efficiency of the general purpose electronic ICT so widely prevalent today is becoming increasingly difficult to improve within the context of its implementation in semiconductor technology. In order to further improve performance and energy efficiency in computation, it may be necessary to invent a new general-purpose architecture and/or implementation technology. The living cell, whose dimensions are only on the order of a few microns, is a powerful information processor that utilizes extremely small amounts of energy (*~10 kT* per bit) and achieves high functional performance. It may be that inspiration can be drawn from the architecture and technologies used by the cell to develop future information processing systems. The cell is a very complex system about which much is yet to be learned but it may provide suggestions for a pathway for more energy-efficient information processing.

Author details

Victor Zhirnov[1], Ralph Cavin[1] and Luca Gammaitoni[2]

1 Semiconductor Research Corporation, USA

2 NiPS Laboratory, Università di Perugia, Italy

References

[1] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices* (1998 Cambridge University Press)

[2] K. Bernstein, R.K. Cavin, W. Porod, A. Seabaugh, J. Welser, "Device and architecture outlook for beyond CMOS switches, *Proc. IEEE* 98 (2010) 2169-2184

[3] J. J. Welser, G. I. Bourianoff, V. V. Zhirnov, R. K. Cavin, "The quest for the next information processing technology", J. Nanoparticle Res. 10 (2008) 1-10

[4] S. Carnot, Reflections on the Motive Power of Heat and on Machines fitted develop this Power, 1824, Translated by R. H. Thruston, ASME (1943).

[5] S. Shankar, R. K. Cavin, and V. V. Zhirnov, "Computation from Devices to System-Level Thermodynamics," *Electrochem. Soc. Transactions* 25 (2009) 421-431

[6] J. von Neumann, Fourth University of Illinois lecture, in *Theory of Self-Reproducing Automata*, A.W. Burks, ed., p. 66. Univ. of Illinois Press, Urbana (1966)

[7] R. Landauer, "Irreversibility and heat generation in the computing process", *IBM J. Res. Dev.* 5 (1961) 183-191

[8] C. H. Bennett, "The thermodynamics of computation - a review", *Int. J. Theoretical Physics* 21 (1982) 905-940

[9] B. Nordman, S. Lanzisera, "Electronics and network energy use: Status and prospects", 2011 IEEE Intern. Conf. Consumer Electron. , pp 245-246

[10] International Technology Roadmap for Semiconductors (ITRS), www.itrs.net

[11] R. K. Cavin, P. Lugli, V. V. Zhirnov, "Science and Engineering Beyond Moore's Law", *Proc. IEEE* 100 (2012) 1720

[12] T. Sekiguchi, K. Ono, A. Kotabe, Y. Yanagawa, "1-Tbyte/s 1-Gbit DRAM architecture using 3-D interconnect for high-throughput computing", IEEE J. Solid-State Circ. 46 (2011) 828-837

[13] W. Gitt, "Information - the 3rd fundamental quantity", *Siemens Review* 56 (1989) 36-41

[14] H. Moravec, "When will computer hardware match the human brain?" *J. Evolution and Technol.* 1 (1998) 1-12

[15] D. E. Dillenberger, D. Gil, S. V. Nitta, M. B. Ritter, "Frontiers of information technology", IBM J. Res. & Dev. 55 (2011) 1:1 - 1:13

[16] Luca Gammaitoni (2012): There's plenty of energy at the bottom (micro and nano scale nonlinear noise harvesting), Contemporary Physics, 53:2, 119-135

[17] C. E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–656, July, October, 1948

[18] L. Gammaitoni, 2011, arXiv:1111.2937; L. Gammaitoni, Nanoenergy Letters, 5, 10, 2013..

[19] J. G. Simmons, "Generalized formula for electric tunnel effect between similar electrodes separated by a thin insulating film", *J. Appl. Phys*. 34 (1963) 1793

[20] L. Gammaitoni, D. Chiuchiù, work in preparation, 2014.

[21] R. U. Ayres, *Information, Entropy and Progress*. New York : AIP Press, 1994.

[22] L. Gammaitoni, "Noise limited computational speed", Applied Physics Letters, 11/2007, Volume 91, p.3, (2007)

[23] V. V. Zhirnov and R. K. Cavin, *Microsystems for Bioelectronics* (Elsevier 2010)

[24] I. P. Batra, "Origin of conductance quantization", Surf. Sci. 395 (1998) 43-45

[25] Y. Imry and R. Landauer, "Conductance viewed as transmission", Rev. Mod. Phys. 71 (1999) S306-S312

[26] V. V. Zhirnov and T. Mikolajick, "Flash Memories", in: "Nanoelectronics and Information Technology", by R. Waser (Ed.) Wiley-VCH 2012, pp. 623-634

[27] V. V. Zhirnov, R. K. Cavin, S. Menzel, E. Linn, S. Schmelzer, D. Bräuhaus, C. Schindler and R. Waser, "Memory Devices: Energy-Space-Time Trade-offs", *Proc. IEEE* 98 (2010) 2185-2200

[28] L. M. Grupp, A. M. Caulfield, J. Coburn, S. Swanson, E. Yaakobi, P. H. Siegel, J. K. Wolf "Characterizing Flash Memory: Anomalies, Observations, and Applications", *MICRO'09*, Dec. 12-16, 2009, New York, NY, USA, p.24-33

[29] N. Derhacobian, S. C. Hollmer, N. Gilbert, M. N. Kozicki, "Power and energy perspectives of nonvolatile memory technologies", Proc. IEEE 98 (2010) 283-298

[30] S. Ji, "The cell as the smallest DNA-based molecular computer", Biosystems 52 (1999) 123-133

[31] R. Sarpeshkar, Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems (Cambridge University Press 2010)

[32] A. Danchin, "Bacteria as computer making computers", *FEMS Microbiol. Rev*. 33 (2009) 3-26

[33] C. T. Fernando, A. M. L. Liekens, L. E. H. Bingle, C. Beck, T. Lenser, D. J. Stekel, J. E. Rowe, "Molecular circuits for associative learning in single-celled organisms", *J. R. Soc. Interface* 6 (2009) 463-469

[34] D. Bray, "Protein molecules as computational elements in living cells", *Nature* 376 (1995) 307-312

[35] N. Ramakrishnan, U. S. Bhalla, and J. J. Tyson, "Computing with proteins", *Computer* 42 (2009) 47-56

[36] L. F. Agnati, D. Guidolin, C. Carone, M. Dam, S. Genedani, K. Fuxe, “Understanding neuronal cellular network architecture”, *Brain Res. Rev.* 58 (2008) 379-399

[37] A. Wagner, “From bit to it: How a complex metabolic network transforms information into living matter”, *BMC Systems Biology* 1 (2007) 33

[38] W. W. Forrest, “Entropy of microbial growth”, *Nature* 225 (1970) 1165-1166

[39] Sungchul Ji, Molecular Theory of the Living Cell: Concepts, Molecular Mechanisms, and Biomedical Applications, Springer, 2012.

[40] P. Hanggi, F. Marchesoni, “Artificial Brownian motors: Controlling transport on the nanoscale”, Rev. Mod. Phys. 81 (2009) 387-442

[41] L. Gammaitoni, P. Hanggi, P. Jung, F. Marchesoni, “Stochastic resonance”, Rev. Mod. Phys. 70 (1998) 223-287

Chapter 8

Electronics for Power and Energy Management

Naser Khosro Pour, François Krummenacher and Maher Kayal

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57281

1. Introduction

There is an increasing demand for energy-efficient wireless sensor networks (WSN) in different sensing and monitoring applications. Many autonomous WSN solutions have been deployed in different areas, including health and lifestyle, automotive, smart buildings, predictive maintenance (e.g., of machines and infrastructure) and active RFID tags [1]. For example, miniaturized wireless body area network (WBAN) platforms that can monitor various biological and physiological signals are highly demanded [2]. These WSN platforms face the main technological challenges of miniaturization, autonomy and manufacturing cost [3]. As long lifetime and small form factor requirements of these applications cannot be provided by miniaturized primary batteries, energy harvesting and charging miniaturized rechargeable batteries could be a good solution for realizing autonomous WSNs. These emerging autonomous ultra-low power (ULP) sensors incorporate energy harvesting source, energy storage device and electronic circuits for power management, sensing and communication into a miniaturized system.

The main energy harvesting sources that have been deployed in WSN applications are vibrational, thermal, photovoltaic, and radio frequency (RF) energy sources. Nanoscale energy sources for solar, kinetic and thermal energy harvesting were thoroughly discussed in the previous chapters. As mentioned before, in a kinetic energy harvester, including electrostatic, piezoelectric, and electromagnetic transducers, motion is transduced to electrical power. These transducers have different power throughputs ranging from almost 1μW/cm^2 up to 100μW/cm^2 depending on environmental conditions [1]. The output voltage depends on the transducer type and typically it tends to be too low in the case of electromagnetic transducers and too high in the case of electrostatic transducers [4]. When two junctions that are made of two dissimilar conductors are kept at different temperatures, an open circuit voltage develops between them based on the seebeck effect that can be harvested by an appropriate thermal energy harvester circuit [5]. Although the source power can reach to 100mW/cm^2 for industrial applications, the temperature gradient and the efficiency are much lower for human body application comparing to machine applications [1]. Ambient RF energy is another source of energy harvesting. Although RF energy harvesting can be very attractive for some applications including RFID tags, for wirelessly powering miniature sensors, either a large area antenna should be used or the RF energy source should be used in close proximity of the sensor node. For example in [6], in order to harvest1.5mW in the target autonomous sensor, a 100 mW RF energy source has to be placed at 20 cm proximity of the sensor. Finally, solar cells are the most widely deployed energy harvesting sources that convert incoming photons into electricity. Solar energy is the most abundant and practical form of ambient energy and miniaturized solar cells are already available in the custom sizes as small as 1mm^2 [7]. Especially for outdoor applications, they are an obvious energy source for autonomous systems, as the source power can reach to 100mW/cm^2. However for Indoor applications, illumination levels are much lower in the order of 100μW/ cm^2. Efficiencies of these energy transducers range from 5–30%, depending on the used material and indoor or outdoor conditions. Thanks to their high efficiencies, solar cells can be good energy sources for miniaturized autonomous wireless sensor nodes, even in the order of a few cubic millimeters size. The energy that harvesting sources provide highly depends on environmental conditions. In many WSN platforms, combinations of different energy harvesting sources have been used to guarantee autonomous operation of the sensor. For example, in [8] both thermal and photovoltaic energy harvesters have been exploited to power an ECG system.

Successful implementation of energy harvesting for WSN applications mainly depends on meeting size, autonomy and cost constraints. Meeting these constraints can be quite challenging and highly depends on the chosen application and data transmission requirements. At present, complex wireless sensor platforms such as iMote [9] have been realized as printed circuit boards (PCB) and cannot be used for realizing millimeter-scale sensors. First of all, bulky batteries are needed to provide the required peak and average power during sensing and data transmission. In addition, most of standard wireless transmission protocols such as Zigbee transceivers [10] require centimeter-scale antennas. In order to replace these bulky batteries by miniaturized energy storage options, not only new materials and architectures such as thin film Li-ION batteries should be developed, but also robust low-power circuits should be developed for sensing and wireless data communication. By using smaller batteries for these low power circuits and developing new wireless transmission methods that require smaller antennas, millimeter-scale wireless sensor will become feasible. These millimeter-scale wireless sensor systems create a lot of new applications, such as continuous monitoring of intraocular pressure (IOP) to detect and track the progression of glaucoma. Current eye pressure measurement techniques are invasive and must be performed at a doctor's office, however by using an implanted sensor, IOP can be logged continuously using a capacitive MEMS sensor and periodically transmitted to a base station to be communicated with the doctor. In [11], an autonomous wireless intraocular pressure monitor microsystem has been presented that incorporates a miniaturized photovoltaic (PV) module, a miniaturized thin film Li-Ion battery and the required electronic circuit for solar energy harvesting, sensor readout

http://dx.doi.org/10.5772/57281

and data communication in a 1.5mm^3 biomedical implant. Since the sensor is implanted in the eye, meeting miniaturization target is crucial; the size of the thin-film battery has been limited to 1mm^2 that has only 1.5uAh capacity.

Although for some WSN applications such as implanted biomedical sensors and monitoring infrastructures, cost is not the main issue, however for most of other applications, current harvesting technologies are still far too expensive. In fact, low sensor cost is vital for the mass deployment of autonomous sensors in smart dust applications. A possible route to realize less expensive sensors is the use of MEMS technology for manufacturing. The devices could then be made on a wafer basis in a batch mode, thereby greatly reducing the size and the cost. But reducing the size of the sensor reduces not only the cost but also the harvested energy. In order to meet stringent power requirements of these sensors, the total power consumption of the WSN platform should be minimized by proper circuit design of all power-hungry components, including wireless transceiver, sensor frontend circuit, processor, and memory blocks. In addition, as miniaturized energy harvesters can supply approximately 10 μW–1 mW, designing efficient power management circuits is also very important and can be quite challenging. In this chapter, we will focus on power management circuits for solar energy harvesting. Another major challenge for realizing millimeter-scale wireless sensors is antenna miniaturization. Low-frequency radios suffer from poor transmission distances while high-frequency radios require high-bandwidth high-power circuits [12].

The system level design of the power management circuit starts by selecting an appropriate energy harvesting source and energy storage device, and the circuit is designed to harvest the maximum energy from the energy harvesting source and store in the energy storage option with maximum efficiency. In fact, as harvested energy is intermittent, either a rechargeable microbattery or a supercapacitor should be used as energy storage device for reliable operation of electronic circuits that are used for power management, sensor interfacing and wireless data transmission. In fact, usually the power that is delivered by the harvester is less than the peak power that the sensor requires during sensing or radio communication. In addition, although the average power that energy harvesting source provides should be higher than the average power consumption of the sensor, a temporary mismatch between power generation and consumption could exist and an energy storage device is needed for continuous operation.

Although the main factor for selecting an appropriate PV module is the power that it can provide in different lighting conditions, other characteristics of the PV module, such as open circuit voltage, short circuit current and maximum power point of the PV module, should be also considered to design a highly efficient solar energy harvester. A PV module provides a DC voltage that highly depends on the illumination level and temperature and can be different from the nominal voltage of the target NiMH or Li-ION battery. In addition, the battery voltage changes from its nominal voltage during battery charging and discharging. As the DC voltage of the PV module differs from the voltage of the target battery, different inductive [13] and switched-capacitor (SC) [14, 15] DC-DC converters have been proposed to harvest energy from a miniaturized PV module to charge a battery. Typically these DC-DC converters either use external capacitors or inductors or they operate at high frequencies and use sophisticated control circuits to reach high efficiency in the fully integrated DC-DC converters. However, by using higher frequencies or deploying complex control circuits, the power consumption of power management circuit Is increased and a high efficiency is not achievable under reduced illumination levels when input power is only a few micro-watts. In the direct charging approach [16], the battery is connected to the PV module through a switch, similar to low dropout regulators. However instead of using error amplifiers [17], low power SC circuits have been used to control the switch. By properly matching the PV module and the target battery, this approach achieves a very high efficiency during charging either NiMH or Li-ION batteries [16, 18]. In addition, as neither inductor nor large capacitors have been used, the die area is much smaller comparing to conventional SC or inductive DC-DC converters.

In addition to charging the target battery, since under reduced light intensity, harvested energy from the PV module may be lower than the power consumption of the microsystem, the energy management circuit should dynamically reduce power consumption of the sensor to avoid complete discharge of the battery. As a result, the microsystem can continue its autonomous operation at a lower speed. The power management circuit can measure the remaining charge of the target rechargeable microbattery or supercapacitor and scales the power consumption of the microsystem up or down according to the energy stored in the battery. In this chapter, first of all, all main blocks of the solar energy harvester, including the power management circuit, energy harvesting source and energy storage device will be discussed thoroughly and after that, the proposed circuit will be used to power a miniaturized autonomous sensor. In fact, size and power requirements of a WSN platform highly depend on the target application and how the measurement results should be processed and communicated with a base station or other target sensors. The target sensor is an autonomous hydrogen gas sensor [19]. In fact, as the use of hydrogen fuels becomes more common, an increased demand for low cost miniaturized hydrogen sensors is expected. Palladium (Pd) nanowire hydrogen sensors, which can be used at room temperature, have good sensitivity thanks to their large surface-to-volume ratio while maintaining low power operation and small form factor [20]. Therefore, these miniaturized sensors are good candidates for ultra-low power (ULP) hydrogen sensing. The readout circuit should measure the change in conductivity of Pd nanowires upon hydrogen exposure [21]. As these nanowires have an undesired thermal cross-sensitivity, temperature is also measured and further compensated during sensor calibration, for accurate measurement of hydrogen concentration [21]. In [39], a miniaturized autonomous sensor has been presented that uses nanowire materials to realize the sensor and the energy harvesting source. The total size of this WSN platform is mainly constrained by the battery and wireless transceiver.

2. Energy storage option

Selecting an appropriate energy storage device is the first step in designing an energy harvester circuit, and the power management circuit should be designed with the target of optimizing the overall power efficiency, based on the selected energy storage device. Millimeter-scale sensors that rely on micro-power energy harvesting, typically have additional stringent constraints on die area and usage of external components. Although the target sensor may consume only a few μA on average for sensing and data transmission, the peak power consumption during wireless data transmission is typically much higher. For example, the TZ1053 wireless transceiver consumes 3.3 mA and 2.8 mA during data transmission and reception [22]. Wireless transceivers that are compatible with standard protocols such as CC2520 that uses Zigbee [23] or CC2570 that uses ANT protocol [24] consume even more during data transmission or reception. Miniaturized supercapacitors, NiMH and Li-ION microbatteries are different solutions that have been used in many wireless sensing platforms. The battery should have enough capacity to provide enough power for a few hours' operation of the sensor, even without energy harvesting. In addition to size and capacity, other characteristics, such as peak discharge current, nominal operating voltage, end-of-charge voltage (V_{EOC}), end-of-discharge voltage (V_{EOD}), cycle life and self-leakage should be considered to design energy-efficient sensors that rely on energy harvesting source. Although tiny supercapacitors, such as GZ115F of CAP-XX [25], can be charged and discharged using much higher currents and have much longer cycle lives in comparison with the rechargeable batteries, they have some disadvantages that make them inappropriate for some WSN platforms including miniaturized autonomous sensors that are normally idle for a very short duty cycle during their operation. The main disadvantage of supercapacitors is their high leakage currents that can be even larger than the current provided by the PV cell under reduced illumination level. In addition, they have much lower energy capacity compared with rechargeable batteries of similar sizes which can provide enough power for continuous operation of the sensor when the energy harvesting source is missing for a few hours. Among lithium-based batteries, state-of-the-art thin film lithium batteries, such as MEC220 of Infinite Power Solutions [26] or CBC050 of Cymbet [27], are a promising technology for integrated energy storage. These microbatteries can be used instead of bulky batteries for millimeter-size WSN platforms as miniaturized batteries are already available in the custom sizes as small as 1mm^2 [27]. These batteries, that have only 200μm thickness, can be used as bare dies to be embedded into modules directly or co-packaged with other integrated circuits. In addition to small size and low thickness, these microbatteries have very low self-leakage, long cycle lives and high discharge currents that make them ideal energy storage options for autonomous wireless sensors with low operation duty cycles. Table 1, compares the main characteristics of this battery with MEC220 thin film lithium battery and GZ115F supercapacitor. This thin film Li-Ion battery supports up to 10000 recharge cycles as can be seen in this table. The charge loss of this battery can be as small as 2% per year in room temperature. Although the capacity of this battery is only 400μAh, it can support a high pulse discharge current of 15mA. The main disadvantage of these batteries is their high nominal voltage levels and their higher cost comparing to conventional NiMH and Li/ION batteries. As thin film lithium batteries have nominal voltages of more than 3.8 V, additional DC-DC converters are required to use these batteries for ULP applications. NiMH microbatteries have a nominal voltage of 1.2 V and can provide relatively high capacity and discharge current. For example, Varta V6HR microbattery [28] has a large nominal capacity of 6.2 mAh and can provide a peak discharge current of 18 mA, while its diameter and height are 6.8 mm and 2.15 mm, respectively. The miniaturized gas sensor that has been proposed in [21] uses this battery as the energy storage option.

Energy Storage Option/ Specification	**GZ115F of CAP-XX [24]**	**MEC220 of Infinite Power Solutions [25]**	**V6HR of Varta [27]**
Technology	Supercapacitor	Thin film battery	NiMH battery
Nominal Voltage	2.3V	4.1V	1.2V
Size	20X15X1.25 mm^3	25.4X25.4X0.17 mm^3	D: 6.8mm, H: 2.15mm
Energy Capacity	4 μAh	400 μAh	6,200 μAh
Peak Discharge Current	30A	15mA	18mA
Cycle life	30,000+ hours	10,000	1,000

Table 1. Comparing energy storage options

Selecting appropriate energy storage option highly depends on the size, cost, wireless communication and autonomy constraints in the target WSN platform. At present, thin film Li-ION batteries can be as small as 1mm^2, but these tiny batteries cannot provide enough discharge current for standard wireless communication protocols. In addition, these batteries are not appropriate for low-cost applications due to their much higher prices comparing to conventional NiMH or Li-ION batteries. If standard wireless communication protocols are supposed to be used in a low-cost application, conventional NiMH or Li-ION microbatteries are still the best options. These batteries should be able to provide required impulse current during data transmission. Finally autonomy requirements should be considered to select a battery with enough energy capacity to provide enough energy during the time interval that energy harvesting source is missing. NiMH microbatteries have been widely used as energy source for different applications and typically have very low prices. As can be seen in this table, Varta V6HR has a relatively high capacity and discharge currents, while its 1.2 V nominal voltage can be an additional advantage in ULP applications. If in the target sensor, all electronic circuits, including the external wireless transceiver, can operate with a sub-1.2 V supply voltage, additional step-down or step-up circuits are not required. The main disadvantage of Varta V6HR NiMH battery in comparison with MEC220 thin film lithium battery is its lower cycle life and higher discharge current. This battery loses almost 20% of its charge during the first month. The main disadvantage of Varta V6HR NiMH battery in comparison with GZ115F is its lower discharge current and shorter cycle life. On the contrary its capacity is much higher and the leakage current is much less than GZ115F. Although supercapacitors have low energy capacity and high leakage current that makes them inappropriate for miniaturized energy harvesting applications where harvested energy is in the order of 10 μW–100 μW, thanks to their high discharge current and their long cycle life, they can be ideal candidates for complex WSN platforms with higher input powers. These supercapacitors can be used as energy buffers besides the rechargeable batteries to provide impulse currents.

3. Solar energy harvester

The solar energy harvester circuit should harvest energy from input solar cells and store it in the target rechargeable battery with the maximum end-to-end efficiency. The main tasks of a solar energy harvester block are battery charging, battery management and energy management. In order to achieve high efficiency, the main characteristics of the PV module, such as open circuit voltage, short circuit current and maximum output power should be considered. A single solar cell can be modeled as in Figure 1 [29]. The open circuit voltage of a miniaturized solar cell is only a few hundred millivolts and if it is supposed to be used to charge a NiMH battery, either inductive boost DC-DC converters or step-up SC DC-DC converters should be used and direct charging approach cannot be deployed. However single solar cells can be connected in series to provide higher DC voltages. In direct charging approach, an appropriate number of solar cells should be used in series to provide sufficient DC voltage. In [21] four miniaturized nanowire solar cells [30] have been connected in series to charge a NiMH microbattery, while to charge a thin film Li-ION battery ten solar cells should be used in a series configuration [18].

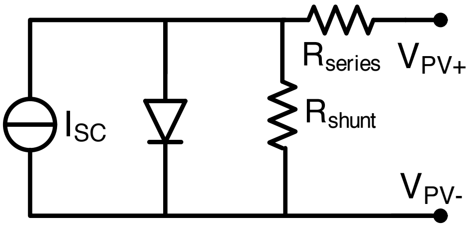


Figure 1. Circuit model of a single solar cell

The block diagram of a miniaturized autonomous sensor that uses direct charging method to charge a NiMH battery is depicted in Figure 2. In the solar energy harvester, the battery voltage (V_{bat}) is compared with the open circuit voltage of the PV module (V_{pv}) using a dynamic comparator, and if it is lower, the battery is connected to the PV module to store harvested energy. If V_{pv} drops below V_{bat}, the switch between V_{pv} and V_{bat} is kept turned off to avoid battery discharge through the PV module. In this approach, the maximum power point voltage (V_{mpp}) of the PV module should be close to the V_{EOC} of the battery to achieve high efficiency. Although P_{mpp} and I_{SC} of the PV module changes considerably in different lightning conditions, V_{oc} and V_{mpp} of the PV module does not change significantly. In Figure 3, the deliverable power of a PV module that uses four solar cells in series has been simulated at different illumination levels using the electrical model of the solar cell [19]. This PV module has a total area of 28 mm^2 and can provide a maximum power P_{mpp} of 2.88 mW in AM1.5 illumination. In AM1.5, solar intensity is almost 1 mW/mm^2. At 10% of AM1.5, although the deliverable power is reduced by roughly the factor of 10, V_{oc} just drops from 2.35 V to 1.95 V.

The second main task of the energy harvester circuit is battery management. In miniaturized sensors that rely on energy harvesting, as the charging current is limited by the PV module, normally the battery is charged in either standard or trickle charging modes, depending on the illumination conditions and capacity of the battery and there is no need to limit charging current [28]. In trickle charging mode, the battery is charged by a small current, and it can be

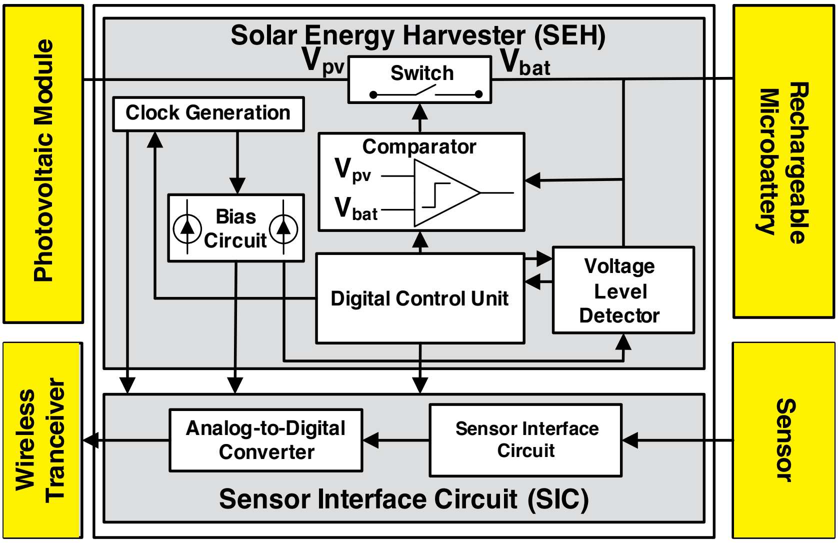


Figure 2. Block diagram of the proposed hydrogen gas sensor

continuously charged, even after reaching V_{EOC}. For example in Varta V6HR, if charging current is less than 3% of its 6mAh energy capacity or equivalently 180μA, the battery is charged in trickle charging mode. Trickle charging limit is normally a function of the energy capacity of the battery and larger batteries have higher charging limits. When battery is charged by a charging current that is higher than this trickle charging limit, the battery charging should be stopped after reaching V_{EOC} to avoid permanent damage to the battery. In addition, to avoid full discharge of the battery that reduces the cycle life of the battery, the battery should be disconnected from the sensor when the battery voltage drops to V_{EOD}. As a result, the battery management unit should detect V_{EOC} and V_{EOD} voltage levels to avoid overdischarge or overcharge of the battery [16]. Finally, the last task of the energy harvester circuit is energy management. Since, under reduced light intensity, the power delivered by the PV module may be lower than the average power consumption of the sensor, the energy harvester circuit should reduce power consumption of the microsystem to avoid full discharge of the battery during this period. Under reduced light intensity, the microsystem continues its autonomous operation at a lower speed and with a lower duty cycle. The energy management circuit can measure the remaining charge of the target battery and scales the power consumption of the microsystem up or down according to the measurement results. In order to reduce the power consumption and speed of digital circuits, the operating frequency is scaled down and the data transceiver is activated by a lower duty cycle. Bias currents are also scaled down to reduce power consumption and the speed of analog circuits. In order to reliably estimate the energy stored in the battery, battery voltage is measured when the battery is discharged by a high discharge current. As harvested solar energy may be as small as a few micro-watts under reduced light intensities, the power consumption of the energy harvester circuit should not be higher than a few hundred nanowatts to have high efficiency during these periods. As neither Vpv nor Vbat changes rapidly, comparator and voltage level detector blocks are activated every few seconds to minimize average power consumption [19].

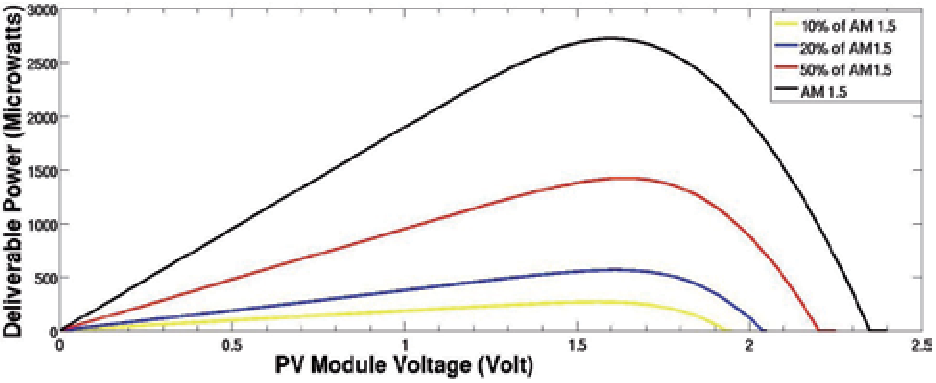


Figure 3. Deliverable power of the PV module in different lightning condition [19]

The dynamic comparator and required control signals can be seen in Figure 4. After turning off the switch between V_{pv} and V_{bat}, the first inverting gain stage gets reset by activating the reset control signal, to establish a common mode voltage of V_{cm} at the V_{C_Top} node. Meanwhile, V_{C_Bot} is connected to V_{pv} by activating the Sample_V_{pv} control signal, and the difference between V_{pv} and V_{cm} is sampled on the sampling capacitor. In the next step, by deactivating the reset and activating Sample_V_{bat} control signal, V_{C_Bot} is connected to V_{bat}, and as a result, V_{C_Top} changes to $V_{cm} + (V_{bat} - V_{pv})$. Outputs of the first and second gain stages (V_{o1} and V_{o2}) change according to the new V_{C_Top}, and V_{o2} is latched and stored as V_{Comp}, by enabling the latch_enable control signal. If V_{Comp} gets 1, it means that V_{bat} is higher than V_{pv}, and the switch should be kept turned off to avoid battery discharge, but if V_{Comp} gets 0, the PV module starts to charge the battery by turning on the switch. After the switch is turned on, V_{pv} will follow V_{bat} during battery charging.

As the operating voltage of the PV module is determined by the battery, end-to-end efficiency from the PV module to the battery is reduced when the battery voltage diverges from the V_{mpp} of the PV module during battery charging. Different maximum power point tracking (MPPT) strategies have been deployed in energy harvesting applications to maximize the amount of harvested energy. As the V_{mpp} of the PV module varies with incident light conditions, an efficient MPPT scheme ensures that the maximum power is extracted from a PV module at any given time. Many complex and accurate MPPT schemes have been investigated for large solar harvesting systems [31]; however, when the PV module is small and can only provide a few micro-watts under reduced light intensity, only low-overhead schemes that incur very little power overhead can be good candidates. Several low-overhead MPPT approaches, including fractional open-circuit voltage (FOC), fractional short circuit current (FSC) and hill-climbing techniques, can be deployed in inductive and SC DC-DC converters [32]. In direct charging V_{pv} always follows V_{bat} and no MPPT scheme can be deployed. Nevertheless, it can achieve even higher end-to-end efficiency than competing SC and inductive DC-DC converters. In [16], end-to-end efficiency from the PV module to the battery has been simulated at different illumina-

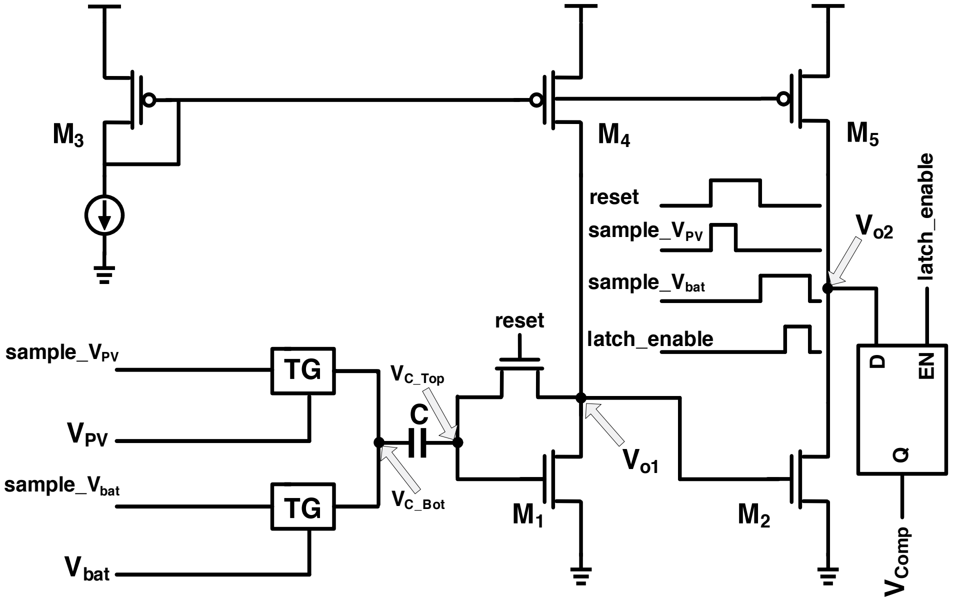


Figure 4. Circuit diagram of dynamic comparator [19]

tion levels. As can be seen in Figure 5a, the miniaturized PV module that has been used as energy harvesting source can provide maximum power of 77.1μW at 1.33V under AM1.5 illumination. The efficiency reaches to the minimum value of 76.9% when the battery is fully discharged. Starting from a fully discharged NiMH battery at V_{POR} and charging it up to fully charged state at V_{EOC}, the average end-to-end efficiency during charging process is 90.6%. In Figure 5b, although illumination is reduced by a factor of 10 and the maximum input power is only 6.36μW, the average efficiency during charging process is still 85.2%. In fact, since the voltage of the PV module does not change significantly at different lighting conditions, the MPP of PV module is still close to nominal voltage of battery and a high end-to-end efficiency from PV module to the battery is achievable without any MPP tracking.

Table 2 compares the simulated performance of the circuit we described with state of the art SC and inductive DC-DC converters that had been previously used for micro power solar energy harvesting. The proposed power management system, achieves better end-to-end efficiency in comparison to these circuits. In addition, as neither inductor nor large capacitors have been used in this architecture, the die area is minimized. Finally as high quality capacitors or inductors are not required, the circuit can be easily realized in different technologies.

When the target rechargeable battery is discharged by a high current, the battery voltage drops and the voltage drop depends on the remaining charge of the battery. For example, when the V6HR NiMH microbattery is fully charged, the battery voltage is close to V_{EOC}; as the battery is discharged by a low current, V_{bat} drops to its nominal value and remains almost constant, up to getting close to the fully discharged state. However, if the battery is discharged by a high current, as can be seen in the discharge curve of the battery in Figure 6 [28], V_{bat} drops

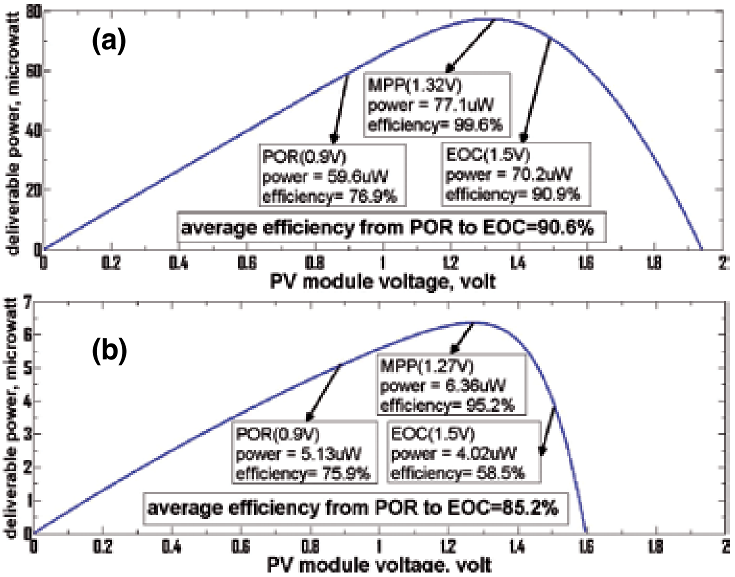


Figure 5. Efficiency of power management system, under simulated AM 1.5 illumination (Figure 5a), under simulated 10% light intensity (Figure 5b)

immediately and voltage drop depends on the remaining charge of the battery. During wireless data transmission, the battery is discharged by a high current, close to 5 CA (C being the 1 hour charge or discharge current). As can be seen in Figure 6, if discharge capacity is less than 20%, V_{bat} is higher than 1.1 V during this period; however, if discharge capacity is higher than 20%, V_{bat} drops from 1.1 V down to 900 mV, depending on the remaining charge of the battery. In our circuit, the battery voltage is detected during this period to accurately estimate the energy stored in the battery. Depending on the detected V_{bat}, the power-performance of the integrated electronic circuits and operation duty cycle of the wireless transceiver are reconfigured to guarantee autonomous operation of the sensor. If V_{bat} is high enough, the sensor can operate at its highest performance: solar energy harvester and sensor interface circuits work at their highest speed, and measurement results are sent to the base station every 10 seconds. If V_{bat} is not high enough, the sensor interface circuit and wireless transceiver are activated with a lower duty cycle to minimize total power consumption of the sensor. If V_{bat} is close to V_{EOD}, these blocks should be deactivated temporarily to avoid full discharge of the battery. Battery voltage can be detected during this period to accurately estimate the energy stored in the battery. Depending on the detected V_{bat}, the power-performance of the integrated electronic circuits and operation duty cycle of the wireless transceiver can be reconfigured to guarantee autonomous operation of the sensor. If V_{bat} is high enough, the sensor can operate at its highest performance: solar energy harvester and sensor interface circuits work at their highest speed, and measurement results are sent to the base station every 10 seconds. If V_{bat} is not high enough, the sensor interface circuit and wireless transceiver are activated with a lower duty cycle to minimize total power consumption of the sensor. If V_{bat} is close to V_{EOD}, these blocks should get deactivated temporarily to avoid full discharge of the battery.

	[13]	[14]	[15]	[16]
process	**0.25µm CMOS**	**0.35µm CMOS**	**0.35µm CMOS**	**0.18µm CMOS**
input voltage (V)	0.5~2	2.1~3.5	1~2.7	0.9~2
output voltage (V)	0~5	3.6~4.4	2	0.9~1.5
power throughput	5µW ~10mW	<780µW	0 ~80µW	0~80µW
controller power	2.4µW ~3.5µW	>10µW	450nW ~850nW	<300nW
end-to-end efficiency	70%	67%	86%	90%

Table 2. Performance comparison between different solar energy harvesters [16]

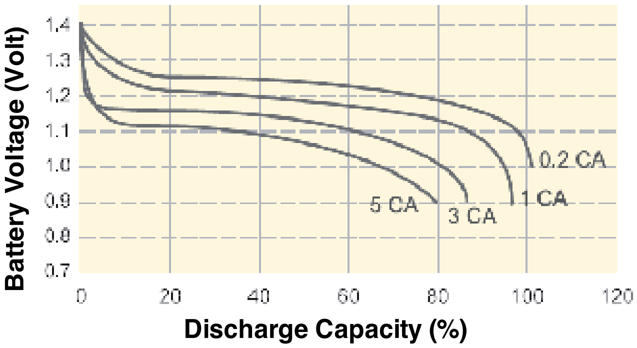


Figure 6. Typical discharge curves of the target NiMH microbattery at room temperature [28]

The voltage level detector (LD) in Figure 7 can be used to determine the battery voltage level [16]. After circuit startup, the battery voltage is checked to make sure that it is more than the end of discharge threshold voltage (V_{EOD}) and can power up the circuit. If V_{bat} is less than V_{EOD}, it means that the battery is fully discharged and cannot provide enough power for the electronic circuits. In this situation, the PV module continuously charges the battery by keeping the switch closed. As soon as V_{bat} passes V_{EOD}, LD starts its normal operation, comparing V_{bat} with specified threshold voltages in a timely manner and updating the V_{bat_level} as a result. In order to generate a bandgap reference voltage, 25 substrate PNP transistors have been used as Q_1 and Q_2 in a common-centroid layout [18]. These transistors are biased with a 100 nA current source to generate V_{BE1} and V_{BE2} in non-overlapping Φ_1 and Φ_2 phases, and the SC circuit sums up V_{BE1} and (V_{BE1}- V_{BE2}) with appropriate coefficients. When V_{bat} reaches V_{EOC}, the switch between the PV module and the battery is turned off to avoid overcharge of the battery.

The SC circuit of Figure 7 operates using non-overlapping clock signals Φ1 and Φ2 and detects when V_{bat} passes the V_L specified in Equation (1) by setting the V_{bat_level} output. Required command signals and status of the switches during Φ_1 phase, can be seen in this figure. In this phase, V_{BE1} is applied to bottom plate of αC and βC capacitors, while bottom plate of γC

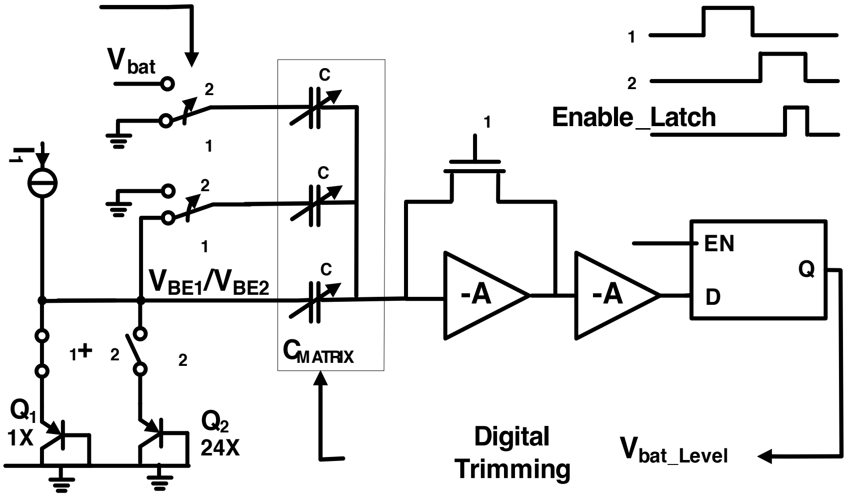


Figure 7. Circuit diagram of voltage level detector [19]

capacitor is grounded. The first inverter gain is reset during this period to establish a fixed common mode voltage for αC, βC and γC capacitors during this period. In phase Φ_2, V_{BE2} is applied to the bottom plate of αC capacitor, while bottom plates of βC and γC capacitors are connected to ground and V_{bat} respectively. As a result the voltage of the top plates of αC, βC and γC capacitors change and this change is amplified by two inverting gain stages and by enabling "Enable_Latch" control signal during phase Φ_2, V_{bat_level} output is updated. By using a variable γC capacitor, different voltage levels between V_{EOD} and V_{EOC} can be detected to estimate the remaining charge of the battery. In Equation (1), α, β and γ coefficients are the ratios of tunable αC, βC and γC capacitors. These variable capacitors have been implemented using a matrix of 10 fF metal-insulator-metal (MIM) capacitors that have been used as unity capacitors.

$$V_L = [\alpha \times (V_{BE1} - V_{BE2}) + \beta \times V_{BE1}] / (\gamma) = V_{ref} / (\gamma) \qquad (1)$$

As V_{BE1} is complementary to absolute temperature (CTAT) and (V_{BE1}- V_{BE2}) is proportional to absolute temperature (PTAT), different CTAT, PTAT or temperature-independent reference voltages (V_{ref}) can be built by proper selection of αC and βC capacitors. After that, by modifying the γC capacitor, target voltage levels can be detected. As can be seen in Table 3, by modifying βC and γC capacitors, different battery voltages, starting from V_{EOD} of 0.9 V up to V_{EOC} of 1.5 V, can be detected. In order to have a bandgap temperature-independent voltage reference, αC should be modified according to the selected βC capacitor. By detecting the battery voltage between 900 mV and 1.1 V, the remaining charge of the battery can be estimated according to the discharge curve of the battery in Figure 6. Similar circuit architecture can be used to detect battery voltage of thin film Li-ION batteries [18]. The same bandgap reference can be used as reference voltage.

Voltage Level (V_L)	αC Capacitor	βC Capacitor	γC Capacitor
907mV(V_{EOD})	170fF	1230fF	230fF
947mV(V_{L0})	170fF	1230fF	220fF
1002mV(V_{L1})	180fF	1300fF	220fF
1051mV(V_{L2})	180fF	1300fF	210fF
1104mV(V_{L3})	190fF	1370fF	210fF
1509mV(V_{EOC})	160fF	1160fF	130fF

Table 3. Detected voltage levels in voltage level detector

After measuring V_{bat} and estimating the energy stored in the battery, the operating frequency of digital circuits and bias currents of analog circuits are reconfigured according to the remaining charge of the battery. In the current-starved ring oscillator in Figure 8a, the frequency is determined by R_L and C_L and can be specified as Equation (2). In Equation (2), K_1 is a constant value that depends on the number of inverter stages in the ring oscillator [33]. The left part of this oscillator is a current mirror that makes the inverter ring biased with a total current identical to the one flowing through RL resistor. This total current is split to charge and discharge all load capacitors. Capacitor C1 is used for decoupling the inverter ring supply voltage and does not participate in the time-constant of the oscillator circuit. By using a digital resistive trimming network to modify R_L, F_{osc} is reconfigured by an energy harvesting circuit, according to the measured V_{bat}.

$$F_{osc} = K_1 / (R_L \times C_L) \tag{2}$$

In addition to a fixed 10 nA beta-multiplier (BM) current reference that has been used to provide the required bias current for a solar energy harvester, a switched-capacitor beta-multiplier (SCBM) current source has been designed to generate frequency-proportional bias currents. In this circuit, which can be seen in Figure 6b, SC resistors have been used instead of regular resistors to generate frequency-proportional bias currents [34]. For example, I_{ADC} can be specified as in (3):

$$I_{ADC} = K_1 \times K_2 \times F_{osc} \times (C_3 + C_4) \tag{3}$$

In Equation (3), K_1 and K_2 are constant values that depend on the ratio between the widths of transistors in Figure 8b. K_1 depends on the ratio between M_7 and M_4, and K_2 depends on the ratio between M_5 and M_6. The current ripple of I_{bias} is minimized by using a large decoupling capacitor, C_2, and two complementary branches charging and discharging C_3 and C_4 capacitors in non-overlapping $Ø_1$ and $Ø_2$ clock phases with F_{osc} frequency [34]. As F_{osc} is determined by the oscillator, I_{bias} scales dynamically by changing the frequency of the oscillator. In addition, this current source can be easily deactivated by turning off the $Ø_1$ and $Ø_2$ clocks.

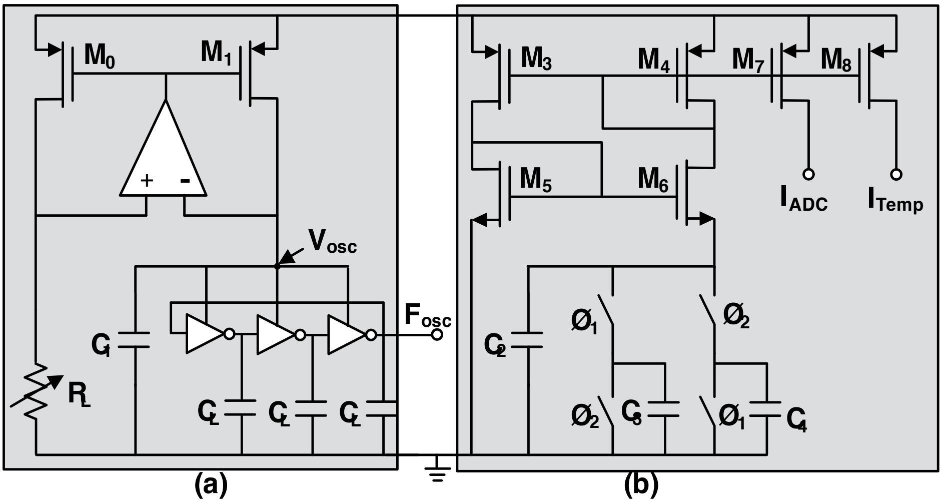


Figure 8. Reconfigurable current-starved ring oscillator (Figure 8a), Switched- capacitor beta-multiplier (SCBM) current source (Figure 8b) [19]

In ultra-low power analog circuits, such as the discrete time incremental ADC that has been used in [21], bias currents should be high enough to guarantee correct operation at the target operating frequency. Bias currents that are generated by SCBM can be used for the sensor interface circuit. If V_{bat} is low, F_{osc} is decreased to reduce the power consumption of digital circuits in the ADC. When the system is operating at a lower frequency, lower bias currents can be used to reduce the power consumption of analog circuits in the sensor interface circuit. By using these frequency-proportional bias currents, power consumption of the sensor can be scaled down dynamically by reducing F_{osc}.

The comparator and LD blocks can be activated in a timely manner, to check the charging status of the battery. As neither V_{bat} nor V_{pv} change rapidly, these blocks are activated every few seconds to minimize their average power consumption. After activating each block, a digital control unit (DCU) generates the required control signals for related SC circuits. The power consumption of DCU is mainly determined by the low frequency counter that activates the comparator and LD in a timely manner. Although comparator and LD blocks consume considerable power during their active time, nevertheless, as they are activated every few seconds, their average power consumption is negligible. The simulated total power consumption of the energy harvester circuit is mainly determined by the clock generator, bias circuit and DCU blocks, which are always active.

4. Wireless transceiver and sensor interface

Emerging wireless sensors that are powered by micro-power energy harvesting sources have more stringent energy requirements compared to traditional wireless sensors. Apart from minimizing the total power consumption of the sensor, which is mainly achieved by duty cycling operation of the sensor, they should have low peak power and ultra-low standby current. Low peak power ensures that miniaturized batteries with limited peak discharge currents can be used to power up the circuit. Ultra-low standby current guarantees that the average power consumption of the sensor can be minimized by heavily duty cycling sensing and data transmission. The main factors impacting power consumption of a wireless transceiver are supply voltage, carrier frequency and receiver sensitivity. The power consumption of transceiver can be reduced by operating at lower supply voltages. Although most of the wireless transceivers work with at least 1.8 V supply voltage, ultra-low power wireless transceivers with sub-1.2 V voltage, such as TZ1053 [22] or ZL70250 [35], are more suited to miniaturized energy harvesting application, thanks to their lower power consumption during data transmission and reception. These transceivers operate at sub-1 GHz frequency bands and have much lower peak-power and standby power, compared to state-of-the-art 2.4 GHz transceivers.

The second important factor is the carrier frequency. Some of the factors that affect choice of carrier frequency are operation range, power consumption, transmission data rates and antenna size. Although 2.4 GHz protocols, such as Zigbee, have been used extensively for wireless sensing applications, sub-1 GHz wireless systems offer several advantages for ultra-low power, low data rate applications. These transceivers have less power consumption for the same operating range, thanks to reduced attenuation rates and blocking effects at lower frequencies. Besides requiring higher power for the same link budget, higher traffic in the 2.4 GHz frequency band increases the interference in this frequency band. Finally, receiver sensitivity also affects the power efficiency. A narrower bandwidth creates higher receiver sensitivity and allows efficient operation at lower transmission rates. Overall, all radio circuits running at higher frequencies, including low-noise amplifiers, power amplifiers, mixers and synthesizers, need more current to achieve the same performance as lower frequencies do. Sub-1 GHz transceivers have some disadvantages, such as larger size of antennas and lower data rates, but overall, they are more suited for our application. A higher data rate can improve the energy efficiency (energy per bit) of the transceiver for high bandwidth applications with large data payloads. In fact, overall power consumption of a wireless transceiver is not only a factor of physical layer items, such as radio architecture, carrier frequency and antenna choice, but is also a function of the amount of time that the radio needs to run in order to transport the payload data over the air. Transmission time depends on the data rate and the protocol overhead to establish and maintain the communication link. For example, a Zigbee transceiver that has a 250 Kbps data rate can send 1 MB of data almost five-times faster than a TZ1053 that has a 50 kbps data rate. However, in ultra-low power applications, typically, data payloads are small and a 50 kbps data rate is more than enough. As standard protocols, e.g., Zigbee or Bluetooth, offer highly sophisticated link and network layers and have a large protocol overhead, they are not very efficient for sending small data payloads.

Miniaturization of WSN platforms highly depends on the chosen application, in addition to using appropriate wireless transceiver. In fact, in some WSN platforms, minimizing the power consumption of the sensor interface circuit can be quite challenging. A miniaturized Pd nanowire grid fabricated on a silicon wafer has been used for hydrogen gas sensing. Each grid consists of 14 Pd nanowires, half of which are covered with a passivation layer to prevent hydrogen from reaching the nanowire [36]. These coated nanowires, which are only sensitive to temperature, have been used as reference nanowires, while the remaining nanowires, which are sensitive to both temperature and H_2 concentration, have been used as sensing nanowires. The sensor interface circuit of Figure 9 measures the conductance change of the sensing nanowires in comparison to the reference nanowires. In addition to eliminating the effects of temperature by first order, the measurement is only sensitive to the ratio of conductance of nanowires, instead of their absolute values, and as a result, a higher accuracy is achievable. Although the temperature effect is eliminated by first order, there are still second order effects, as the temperature coefficient of nanowire resistance changes according to H_2 concentration. In order to compensate this second order effect, temperature is measured using an integrated temperature sensor. H_2 sensing accuracy can be increased by incorporating the measured temperature during sensor calibration. An incremental analog to digital converter (ADC) [37] converts the measured temperature and H_2 concentration to 12-bit digital values at different times. Individual Pd nanowires that have between 7 KΩ and 9 KΩ resistance should be biased, with a minimum bias voltage of 50 mV for 10 seconds. These nanowires have been represented by NW_{ref} and NW_{sense} in Figure 9. The voltage around the reference nanowire, $V_R = (V_1 - V_2)$, is used as the reference voltage, while the voltage of the sensing nanowire (V_2) is used as the input voltage for the following incremental ADC in consecutive non-overlapping clock phases. As a 1MHz clock has been used for ADC and since this ADC needs 2^N cycles for N-bits conversion, a 12-bit conversion takes nearly 4 milliseconds. After gas sensing is completed, temperature is measured to further increase the accuracy of gas sensing using an integrated temperature sensor. In the proposed integrated temperature sensor, substrate PNP transistors have been used to generate proportional to absolute temperature (V_{ptat}) and temperature-independent reference (V_{ref}) voltages, and the ADC converts (V_{ptat}/V_{ref}) to a 12-bit digital value. In Figure 9, by using a 50 fF MIM capacitor as C_1 and a 360 fF MIM capacitor as C_2, a temperature-independent reference voltage of approximately 310 mV has been generated. Similar to gas sensing, n-bit digital representation of (V_{ptat}/V_{ref}) is stored in the ADC counter and sent to a wireless transceiver after temperature sensing. The accuracy of temperature sensing is mainly limited by mismatch between Q_1 and Q_2 and nonlinearity in temperature dependence of V_{BE1} and (V_{BE1}- V_{BE2}). Although these errors can be minimized by using dynamic methods presented in [38] to reach ±0.1°C accuracy, such power-consuming techniques are not needed here. The proposed low power temperature sensor can achieve ±1 °C accuracy by only calibrating C_2 and I_{Temp} at room temperature. When the sensor interface circuit operates at a lower frequency, Q_1 and Q_2 transistors are biased with a lower I_{Temp} bias current to reduce the average power consumption of the circuit. Power consumption and conversion time of this ADC can be reconfigured according to the energy stored in the battery to match the amount of available power, resulting in an adaptive, autonomous sensor. The power management circuit dynamically reduces the operating frequency of digital circuits and the bias currents of analog circuits in this sensor interface circuit under reduced light intensity. However, ADC conversion time increases as a result.

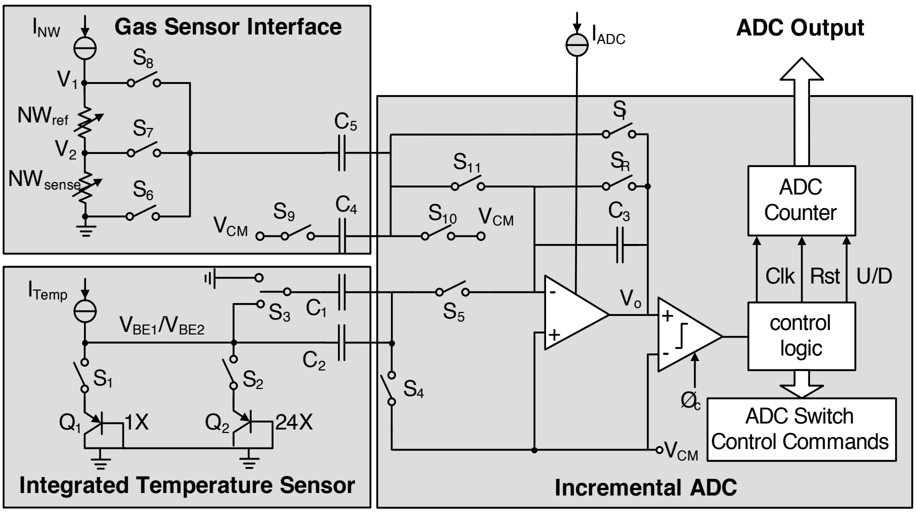


Figure 9. Circuit diagram of the proposed sensor interface [19]

5. Experimental results

The power management and sensor interface circuits have been implemented in a 0.18 μm CMOS process with 0.25 mm^2 total area, as can be seen in the chip microphotograph in Figure 10. The main blocks, including the digital and analog blocks of solar energy harvester and sensor interface circuits, have been specified separately. When operating at 1 MHz, the whole circuit consumes almost 2.1 μW; while the average power consumption of the energy harvester is less than 350 nW. The wireless transceiver sends the measurement results for both temperature and H_2 concentration to a base station every 15 seconds.

Table 4 presents measurement and simulation results for power consumption of the energy harvester and sensor interface circuits in different system operation modes. These operation modes have been defined according to the remaining charge of the battery, which is estimated accurately by detecting the battery voltage between 900 mV and 1.1 V during wireless data transmission, and the battery is discharged by a high current, close to 5 CA. Battery discharge capacity is determined according to the discharge curve of the battery shown in Figure 6. Highly resistive poly resistors have been used in the beta multiplier current source, and oscillator. As these embedded resistors are sensitive to process variations, they have been trimmed initially. If the battery gets discharged, a lower battery voltage is detected during 5 CA discharge and the system is switched to a lower clock frequency to decrease the average power consumption of the whole system. In S_{L2} and S_{L1} modes, the operating frequency is decreased to 500 KHz and 250 KHz, respectively. In S_{L0} mode, when the remaining charge of the battery is less than 25%, the circuit consumes less than 0.6 μW by operating at 125 KHz frequency instead of 1 MHz and using lower bias currents. The average power consumption of the energy harvester drops to less than 110 nW in S_{L0} mode. In addition, measurement results are sent every 120 seconds instead of every 15 seconds in S_{L3} mode, to further reduce the total average power consumption of the whole sensor.

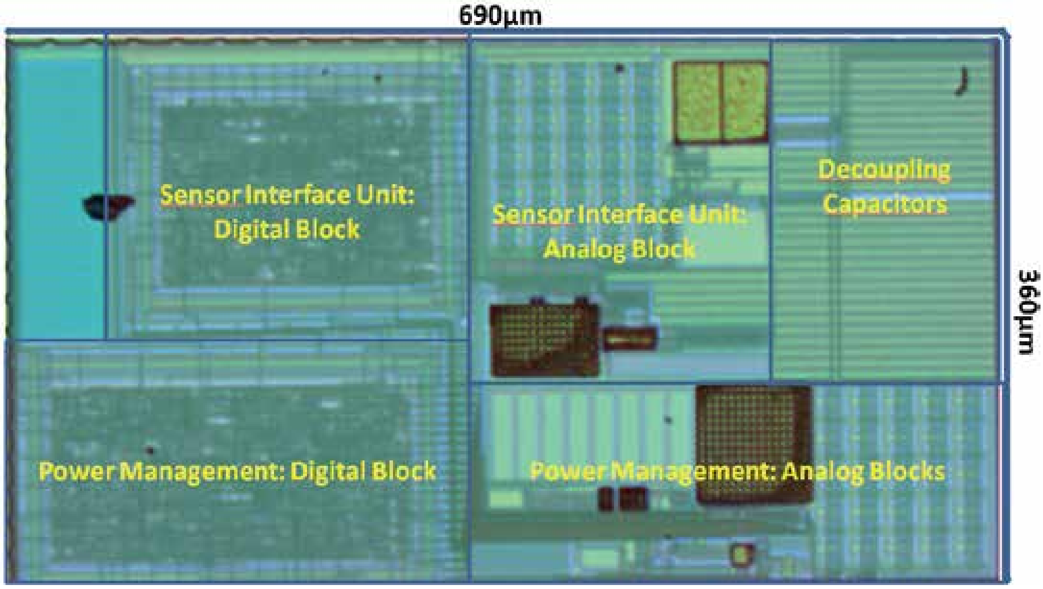

690µm
360µm
Sensor Interface Unit: Digital Block
Sensor Interface Unit: Analog Block
Decoupling Capacitors
Power Management: Digital Block
Power Management: Analog Blocks

Figure 10. Chip microphotograph of the whole microsystem

Block		S_{L3}	S_{L2}	S_{L1}	S_{L0}
Detected V_{bat} during 5CA discharge (mV)		1104	1051	1002	947
Battery discharge capacity (%)		<20%	<50%	<65%	<75%
Battery threshold voltage (Volt)	Measured	1.114	1.055	1.003	0.955
	Simulated	1.122	1.063	1.010	0.962
Clock frequency (KHz)	Measured	950	488	249	127
	Simulated	980	502	257	132
Time interval of sensing and data transmission (Seconds)		15	30	60	120
Power consumption of clock generator (nW)		165	85	44	23
Power consumption of digital control unit (nW)		90	46	24	13
Average power consumption of energy harvester (nW)	Measured	346	210	142	103
	Simulated	293	169	106	74
Average power consumption of sensor interface circuit (nW)	Measured	1730	1120	790	640
	Simulated	1360	870	625	500
Average power consumption of the integrated circuit (nW)		2076	1330	932	707
Average current consumption of gas sensor bias circuit (µA)		4.67	2.33	1.16	0.58
Average current consumption of the wireless transciever (µA)		9.4	7.7	6.6	5.55
Average current consumption of the complete system (µA)		16	11	8.5	6.7

Table 4. System performance and power consumption in different system operation modes

In order to estimate total power consumption of the sensor, the average power consumption of sensor biasing circuit and wireless transceiver should be calculated. Pd nanowires should be biased with a 7 μA bias current for 10 seconds, before measuring H_2 concentration. TZ1053 consumes 5 μA during standby and consumes 3.3 mA during a period of 20 ms to send a sample with the minimum payload size of 55 bytes [21]. In S_{L3} mode, samples are sent every 15 seconds, and the average current consumption of Pd nanowire sensors and wireless transmission are 4.67 μA and 9.4 μA, respectively. By sending the samples every 120 seconds in S_{L0} mode, these values will be reduced to 0.58 μA and 5.6 μA, respectively. So, the total average current consumption of the whole sensor is less than 16 μA, operating at its highest performance in S_{L3} mode, and is reduced to less than 7 μA, operating in S_{L0} mode.

7. Conclusion

In this chapter challenges in designing energy harvester circuits for miniaturized sensing applications were discussed. Successful implementation of energy harvesting for wireless sensing applications mainly depends on meeting size, autonomy and cost constraints. Miniaturization of WSN platforms can be quite challenging and highly depends on the chosen application and data rate of the transmitted data. Typically the total size of a wireless sensing platform is mainly constrained by the battery and wireless transceiver. In this chapter, system level design challenges, including the selection of appropriate energy storage device and wireless transceiver were discussed. Direct charging approach was presented as an ultra-low power energy-efficient solar energy harvesting approach and a sensing microsystem was also presented for wireless sensing applications.

Author details

Naser Khosro Pour*, François Krummenacher and Maher Kayal

*Address all correspondence to: naser.khosropour@epfl.ch

Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

References

[1] Vullers, R.J.M.; Schaijk, R.V.; Visser, H.J.; Penders, J.; Hoof, C.V.;, "Energy Harvesting for Autonomous Wireless Sensor Networks," Solid-State Circuits Magazine, IEEE, vol.2, no.2, pp.29-38, Spring 2010

[2] G.-Z. Yang, Ed., Body Sensor Networks. London: Springer-Verlag, 2006

[3] J. Penders, B. Gyselinckx, R. Vullers, M. De Nil, S. Nimmala, J. van de Molengraft, R. Yazicioglu, T. Torfs, V. Leonov, P. Merken, and C. Van Hoof, “Human++: From technology to emerging health monitoring concepts,” in Proc. 5th Int. Workshop Wearable and Implantable Body Sensor Networks, Chin. Univ. Hong Kong, China, June 1–3, 2008, pp. 94–98.

[4] T. Sterken, K. Baert, C. Van Hoof, R. Puers, G. Borghs, and P. Fiorini, “Comparative modeling for vibration harvesters,” in Proc. IEEE Sensors Conf., 2004, pp.1249–1252.

[5] Lhermet, H.; Condemine, C.; Plissonnier, M.; Salot, R.; Audebert, P.; Rosset, M.;, "Efficient Power Management Circuit: From Thermal Energy Harvesting to Above-IC Microbattery Energy Storage," Solid-State Circuits, IEEE Journal of, vol.43, no.1, pp. 246-255, Jan. 2008

[6] R. J. M. Vullers, H. J. Visser, B. Op het Veld, and V. Pop, “RF harvesting using antenna structures on foil,” in Proc. PowerMEMS 2008, Sendai, Japan, Nov. 10–11, 2008, pp. 209–212.

[7] B.W. Cook, S. Lanzisera, K.S.J. Pister, "SoC Issues for RF Smart Dust," Proceedings of the IEEE, vol.94, no.6, pp.1177-1196, June 2006

[8] V. Leonov, T. Torfs, I. Doms, R. F. Yazicioglu, Z. Wang, C. Van Hoof, and R. J. M. Vullers,“Wireless body-powered electrocardiography shirt,”in Proc. Smart Systems Integration, Brussels, Apr. 10–11, 2009, pp. 307–314.

[9] Datasheet Imote 2 (2009). High-performance Wireless Sensor Network Node. [Online]. Available: www.xbow.com

[10] D. Gislason, ZigBee Wireless Networking. London: Newnes Publications, 2008.

[11] Chen, G.; Ghaed, H.; Haque, R.; Wieckowski, M.; Yejoong K.; Gyouho K.; Fick, D.; Daeyeon K.; Mingoo S.; Wise, K.; Blaauw, D.; Sylvester, D. A cubic-millimeter energy-autonomous wireless intraocular pressure monitor. ISSCC, pp.310-312, Feb. 2011

[12] Chen, G.; Hanson, S.; Blaauw, D.; Sylvester, D.;, "Circuit Design Advances for Wireless Sensing Applications," Proceedings of the IEEE, vol.98, no.11, pp.1808-1827, Nov. 2010

[13] Qiu, Y., Liempd, C.V., Veld, B.O.H., Blanken, P.G., Hoof, C.V.: ‘5μW-to-10mW input power range inductive boost converter for indoor photovoltaic energy harvesting with integrated maximum power point tracking algorithm’, ISSCC, pp.118-120, Feb. 2011

[14] Hui, S., Chi-Ying, T., Wing-Hung, K.: ‘The Design of a Micro Power Management System for Applications Using Photovoltaic Cells With the Maximum Output Power Control’, IEEE Tran. VLSI Systems, vol.17, no.8, pp.1138-1142, Aug. 2009

[15] Jungmoon, K., Jihwan, K., Chulwoo, K.: ‘A Regulated Charge Pump With a Low-Power Integrated Optimum Power Point Tracking Algorithm for Indoor Solar Ener-

gy Harvesting', IEEE Tran. Circuits and Systems II: Express Briefs, vol.58, no.12, pp. 802-806, Dec. 2011

[16] Khosro Pour, N.; Krummenacher, F.; Kayal. M. Fully integrated ultra-low power management system for micro-power solar energy harvesting applications. Electronics Letters, vol. 48, p. 338-U118, 2012.

[17] M. Kayal, F. Vaucher and Phi. Deval, "New Error Amplifier Topology for Low Dropout Voltage Regulators Using Compound OTA-OPAMP," Proceedings of the 32nd European Solid-State Circuits Conference, ESSCIRC 2006, pp.536-539, Sept. 2006

[18] Khosro Pour, Naser; Facchin, Stefano; Krummenacher, Francois; Kayal, Maher;, "An ultra-low power li-ion battery charger for micro-power solar energy harvesting applications," Electronics, Circuits and Systems (ICECS), 2012 19th IEEE International Conference on, vol., no., pp.516-519, 9-12 Dec. 2012

[19] Pour, N.K.; Krummenacher, F.; Kayal, M. Fully Integrated Solar Energy Harvester and Sensor Interface Circuits for Energy-Efficient Wireless Sensing Applications. J. Low Power Electron. Appl. 2013, 3, 9-24.

[20] Offermans, P.; Tong, H. D.; Van Rijn, C. J. M.; Merken, P.; Brongersma, S. H.; Crego-Calama, M. Ultralow-power hydrogen sensing with single palladium nanowires. Applied Physics Letters, vol.94, no.22, pp.223110-223110-3, Jun 2009

[21] N. Khosro Pour, F. Krummenacher and M. Kayal. A miniaturized autonomous microsystem for hydrogen gas sensing applications, New Circuits and Systems Conference (NEWCAS), 2012 IEEE 10th International, vol., no., pp.201-204, 17-20 June 2012

[22] TZ1053 Datasheet [Online] Available: www.toumaz.com

[23] CC2520 Datasheet [Online] Available: www.ti.com

[24] CC2570 Datasheet [Online] Available: www.ti.com

[25] GZ115F Datasheet [Online]. Available: http://www.cap-xx.com

[26] MEC220 Datasheet [Online]. Available: http://www.infinitepowersolutions.com

[27] CBC050 Datasheet [Online]. Available: http://www.cymbet.com

[28] Varta V6HR Datasheet. [Online]. Available: http://www.varta-microbattery.com

[29] Chao, Lu; Raghunathan, V.; Roy, K. Micro-scale energy harvesting: A system design perspective, Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific, vol., no., pp.89-94, 18-21 Jan. 2010

[30] Jia, G.; Steglich, M.; Sill, I.; Falk F. Core–shell heterojunction solar cells on silicon nanowire arrays. Solar Energy Materials and Solar Cells, Vol.96, Jan. 2012, Pages 226-230, ISSN 0927-0248

[31] Esram, E.; Chapman, P. L. Comparison of photovoltaic array maximum power point tracking techniques. IEEE Transactions on Energy Conversion, vol. 22, No. 2, pp. 439-449, June, 2007

[32] Chao, Lu; Raghunathan, V.; Roy, K. Maximum power point considerations in micro-scale solar energy harvesting systems. Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, vol., no., pp.273-276, May 30 2010-June 2 2010

[33] Pastre, M.; Krummenacher, F.; Kazanc, O.; Khosro Pour, N.; Pace, C.; Rigert, S.; Kayal, M. A solar battery charger with maximum power point tracking. 18th IEEE International Conference on Electronics, Circuits and Systems (ICECS) 2011, pp.394-397, 11-14 Dec. 2011

[34] Pastre, M.; Krummenacher, F.; Robortella, R.; Simon-Vermot, R.; Kayal, M. A fully integrated solar battery charger. Joint IEEE North-East Workshop on Circuits and Systems and TAISA Conference, NEWCAS-TAISA 2009, pp.1-4, June 28 2009-July 1 2009

[35] Zarlink ZL70250 Datasheet. [Online]. Available: http://www.zarlink.com/zarlink

[36] Van der Bent, J.F.; Van Rijn, C.J.M. Ultra low power temperature compensation method for palladium nanowire grid. Procedia Engineering, vol. 5, 2010, pp. 184-187

[37] Pertijs, M.A.P.; Makinwa, K.A.A.; Huijsing J.H. A CMOS smart temperature sensor with a 3σ inaccuracy of ±0.1°C from -55°C to 125°C. IEEE Journal of Solid-State Circuits, vol.40, no.12, pp. 2805- 2815, Dec. 2005

[38] Markus, J.; Silva, J.; Temes, G.C. Theory and applications of incremental ΔΣ converters. IEEE Transactions on Circuits and Systems I: Regular Papers, vol.51, no.4, pp. 678- 690, April 2004

[39] Koshro-Pour, N.; Kayal, M.; Jia, G.; Eisenhawer, B.; Falk, F.; Nightingale, A.; De Mello, J.C; et al. "A miniaturised autonomous sensor based on nanowire materials platform: the SiNAPS mote." In SPIE Microtechnologies, pp. 87631Q-87631Q. International Society for Optics and Photonics, 2013.

Chapter 9

Power consumption Assessment in Wireless Sensor Networks

Antonio Moschitta and Igor Neri

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57201

1. Introduction

Wireless Sensor Networks (WSNs) are an emerging technology with a wide range of potential applications. A large number of nodes, with sensing and wireless communications capabilities, deployed in an area of interest, build a WSN. Thanks to the advances in MEMS (Micro Electronics Mechanical Systems) it is nowadays possible to realize small and cheap devices, capable of wireless communication. WSNs differ from other wireless technologies because of a set of specific requirements and characteristic features, including for instance node density, energy requirements, and computing capabilities. The Institute of Electrical and Electronics Engineers (IEEE) classify network technologies by such characteristics. Usually the WSNs are limited to 1Mbps of data rate and 1km of wireless coverage. The actual limit of such quantities depends on the adopted technologies and constraints introduced by specific applications. A set of WSNs specifications, dealing with both network operation and node architectures is described in the IEEE Standard 802.15 and 1451 family. [1-2].

An additional parameter is the WSN operational life, which strongly depends on the balance between power consumption and energy storage. In particular, WSNs are characterized by limited power storage, with possible mitigation coming from power harvesting techniques. Nevertheless, energy efficiency is a critical issue, to be pursued both at node and network level. Typical assumptions include considering the radio interface as the main contributor to power consumption. As a consequence, great attention has been given in the literature to protocol optimization, aimed for instance at minimizing the amount of data transmissions throughout the network, and the maximization of node low-power residence time [3-5]. However, designing a sustainable WSN, relying on power harvesting techniques, would require a deeper and careful modeling, due to the limited and non-steady power supply achievable through harvesting techniques, establishing for instance the maximum consented duty cycle for each node. Such a scenario may require modeling and accurately measuring power consumption associated to the activation of other node functional blocks, in addition to the well-known RF interface. To this aim, both simulation techniques and measurement procedures can be found in the literature. Simulation techniques are available both to describe the network behavior and the node behavior, the latter being based on code profiling techniques and on the description of the node as a finite state machine [6]. Measurement procedures are described as well, typically relying on current measurements at the power supply input, assuming a constant supply voltage [7]. This kind of measurement needs to satisfy conflicting constraints, since it requires to accurately measure short phenomena occurring at a low rate. Moreover, in a distributed context, timing information should be provided, since providing spatial-temporal coordinates to energy consumption measurement may help characterizing the network activity and its operational life.

Following such ideas, the rest of the chapter is organized in three main sections. In section 2, the main features of a WSN are recalled, describing the node architecture and the most popular network topologies, protocols and reference standards. In section 3 energy awareness problems are highlighted, while in section 4 the main techniques for assessing node and network power consumption are recalled, considering both simulators tools and measurement procedures. In the final section, a case study is presented, demonstrating some of the presented approaches.

2. Wireless sensor node and network architecture

2.1. Sensor node structure

Several units compose each node of the WSN, as represented in Figure 1. The core of the wireless sensor node is the processing unit, usually a microprocessor with a limited amount of memory. The processing unit is connected to the sensors via one or more Analog to Digital Converters (ADCs). The sensors and the ADCs form the sensing unit. The data received by the sensing unit are processed and eventually transmitted by the transceiver unit. The transceiver unit is usually capable of bidirectional communications; nevertheless specific applications may require only transmission (TX) or reception (RX) capabilities. Specific nodes may integrate a location finding system that helps the node to discover its position, relative to its neighbors or global. This unit is often embedded on the transceiver module and requires the use of specific algorithms by the processing unit, depending on the adopted localization techniques [8-9].

The power unit and the power generator are a key element in the sensor structure. The power unit is responsible to provide the electrical power needed by the other units in the system. Smart power units are also capable to provide information on the residual available energy, in order to apply energy aware decisions and consent the processing unit to complete the task at hand. Since the power generator usually consists of batteries, such devices have limited amount of energy available, thereby limiting the lifetime of the node. In recent year there has been a big effort in finding alternative solution to power such nodes using the energy available on the node environment with good results [10-12].

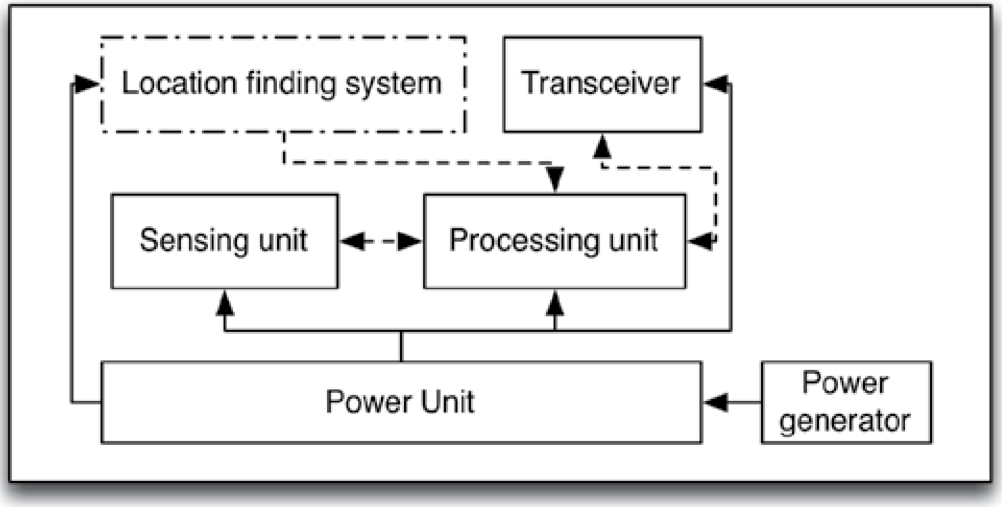


Figure 1. Sensor node architecture

While several off-the-shelf WSN platforms are available on the market, no one of them is considered as a standard de facto. Each research group or industry tends to realize its own platform depending on the objective. This is mostly due to the high cost of existing solutions compared to the costs of the components. Table 1 shows a list of the most popular wireless sensor nodes with their characteristics in terms of computational capacity and radio specifications.

Moreover, it is well known that one of the major constraints in WSN nodes is the low computational capability [13]. Conventional operating systems run on 32/64-bit microprocessors with hundreds or thousands MHz and several MB or GB of memory. For this reason in wireless nodes the applications are specifically designed for the hardware or rely on tailored version of operating systems (OSs), specifically designed for WSN.

Node	CPU	Radio
MICA2	ATmega128L - 16bit 8MHz	868/916 MHz
Telos	TI MSP430 - 16bit 8MHz	2.4GHz IEEE 802.15.4
MICAz	ATmega128L - 16bit 8MHz	2.4 GHz IEEE 802.15.4
ez430-RF2500	TI MSP430 - 16bit 16MHz	2.4 GHz SimpliciTI
WSN430	TI MSP430 - 16bit 8MHz	ISM band (315 to 915 MHz)
VirtualSense	TI MSP430 - 16-bit 25MHz	2.4 GHz IEEE 802.15.4
Pinoccio	ATmega128RFA1 - 8bit 16MHz	2.4 GHz IEEE 802.15.4

Table 1. List of some popular wireless sensors nodes within their characteristics in terms of computational capacity and radio specifications.

Since WSNs can be used to monitor mission critical systems, a Real-Time Operating System (RTOS) is often required. However, only few of the most adopted OSs support RT-applications. According to Farooq and Kunz in [14] the most popular OSs in WSNs are: TinyOS, Contiki, MANTIS, Nano-RK and LiteOS. It is worthy to say that the adoption of an OS increases the power consumption of the node, introducing an overhead due to the management of the process scheduling; however this overhead in power consumption may be compensated by an increase in flexibility in application development. Since different application scenarios call for different tradeoffs, a case-by-case evaluation is required.

2.2. Network architecture

A WSN is usually composed by a large number of nodes deployed in a region of interest. In a typical scenario the region of interest is often a harsh environment, and the nodes are randomly deployed. The sensed data are transmitted through the nodes up to special entities called sinks. The sinks are nodes with two or more network interfaces that act as gateway between the WSN and the user network (e.g. a LAN, or the Internet). The sink usually collects and processes the data from the network sending only relevant information to the user. It also receives commands from the user to be executed on the internal network. A sensor node can communicate directly to the sink (single-hop) or use a multi-hop communication passing the information to its neighbor. Single-hop communication leads to long distance transmission, resulting in high-energy consumption. Using multi-hop communication it is possible to reduce the transmission distance increasing the network lifetime. In multi-hop transmission the network architecture plays a major role. Multi-hop network architectures are typically divided in flat or hierarchical as represented in Figure 2.

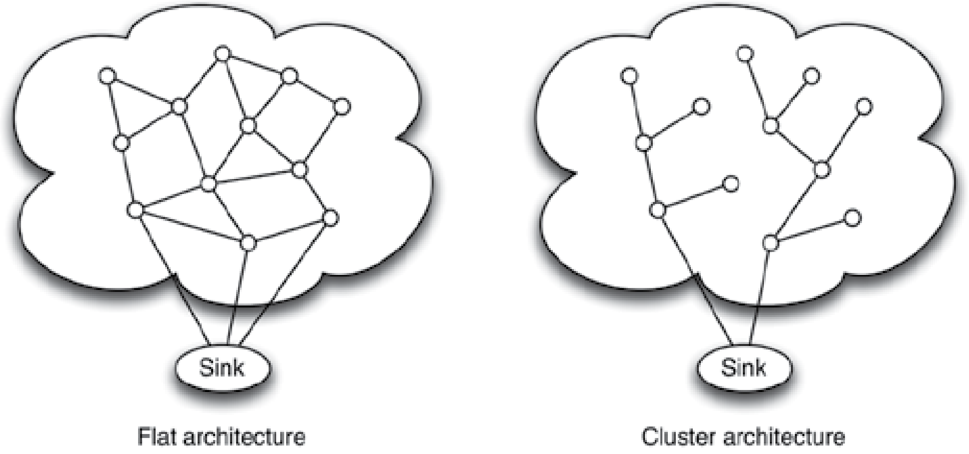


Figure 2. Network architecture

In flat architecture each node plays the same role in sensing and transmitting the information. In hierarchical architectures the nodes are organized into clusters. In each cluster, one or more nodes (head node) are responsible to communicate with other clusters or directly to the sink. The head node may be dynamically selected by various criteria, including its available energy, distance between cluster members and other cluster heads [15], and node homogeneity [16-17].

2.3. Wireless sensor network standards

Interoperability between different products is guaranteed by compliance to standards. In this way different sensors, produced by different manufacturers can communicate and achieve a common objective in the WSN. During the last years different WSN standards have been developed, taking into account different objectives, levels of abstraction and node components. IEEE defines two standard related to WSNs, the IEEE 802.15 regarding the wireless communication interface between nodes and the IEEE 1451 that defines the interface between sensors and actuators. Among the several subgroups of the IEEE 802.15 family the most important, for the WSNs is the IEEE 802.15.4, which specifies the physical layer and media access control (MAC) for low-rate wireless personal area networks (LR-WPANs). The standards support different frequency bands, number of channels and data rates as represented in Table 1.

Band	868 MHz	915 MHz	2.4 GHz
Region	EU, Japan	US	Worldwide
Channels	1	10	16
Data rate	20 kbps	40 kbps	250 kbps

Table 2. Frequency bands, number of channels and data rates for the IEEE 802.15.4 standard.

The preferred frequency band for WSNs is the 2.4 GHz since it is worldwide usable and it has the highest data rate. Moreover the high radio data rate reduces frame transmission time, reducing the microcontroller idle time, resulting in overall energy consumption reduction.

Based on the IEEE 802.15.4 standard several specifications have been developed by different consortia. One of the most famous is ZigBee, a protocol that defines the network and the application layers, built upon the IEEE 802.15.4 physical and MAC layers. Following the ZigBee specification the network layer provides support to tree, star, point-to-point (mesh) network topologies using three different kinds of nodes:

- Network coordinator: it forms the root of the network tree and might bridge to other networks.
- Router: it can act as an intermediate router, passing on data from other devices.
- End device: it cannot relay data from other devices; it is only able to communicate with its parent (router or coordinator).

The application layer provides a framework for distributed application development and communication. On the application layer it is possible to develop up to 240 application objects, which are user defined application modules implementing a ZigBee application. In this way several different applications can reside on a single node, sharing the lower stack of the protocol. Each application object in the network is identified by the network address of the hosting device and the application endpoint number (from 1 to 240).

3. Energy awareness and control of power consumption

The number of existing and prospecting applications has been steadily growing after the development of the WSN paradigm. Regrettably, the energy density of the batteries did not follow the same trend, and the energy harvesting systems can power only a limited class of devices, usually with limited capabilities [10]. For this reason energy consumption modeling and reduction has attracted the interest of both the academic and the industrial worlds. The next sections are devoted to the exploration, modeling, characterization and analysis of the power consumption of a WSN node in relation to specific application. In this section a brief and non-exhaustive review of methods to reduce the power consumption of the nodes is presented.

Due to the limited computational capabilities of the WSN node its load is often limited to trivial computation. The greatest part of energy is spent by the peripherals, especially by the radio module. Thus, a lot of power-saving mechanisms exploit the energy consumption reduction of the node peripherals. In this regard, both passive and active approaches are possible. Passive power conservation mechanisms reduce the energy consumption of a sensor node by turning-off its transceiver interface module when there is no communication activity [18]. Moreover, additional energy savings may also be achieved by optimizing the performance of the processor in an active state changing its operational frequency [19]. In fact, using a processing unit with variable processors speed (VPS), it is possible to decrease its power consumption decreasing the supply voltage and the clock frequency. Exploiting the VPS it is desirable to design a scheduling system, capable to select a suitable supply voltage and relative frequency clock for each task. Dynamic Voltage Scheduling (DVS) is one of these mechanisms able to provide such behavior without degrading the overall performance of the node [19]. Dynamic Power Management (DPM) is another technique to increase the lifetime of a sensor node [20]. DPM acts similarly to DVS, but instead of scaling the clock frequency it can dynamically turn-off the components of the sensor node and wake them up when needed. At microcontroller level this transition of states it is represented by different power mode that shutdown the CPU, memory or additional internal peripherals. It is worth to say that each transition of state takes a certain amount of time and consequent energy consumption as reported in Figure 3. In each power mode, also called low power mode (LMP), different peripherals are incrementally turned off. Each transition from the idle state to a LPM has a fixed cost, indicated in Figure 3 as *b0*, which is usually negligible. However the energy cost for waking up the microcontroller from a low power mode increases with the depth of the low power modes. For this reason it is important to reduce the number of state transitions, conveniently balancing the scheduling mechanism without using aggressive power down strategies.

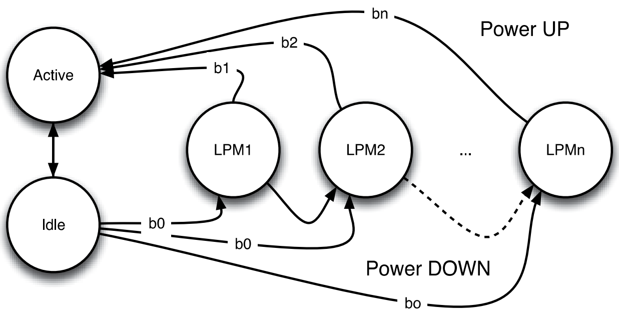


Figure 3. Low Power Modes (LPMs) transitions and costs

Active power conservation mechanisms differ from passive ones in that they achieve a reduction of the energy consumption by avoiding undesired events like collisions, or exploiting energy-aware routing protocols. For instance adjusting the transmission power may help minimizing the probability of occurrence of a collision, an event leading to higher power consumption due to the related detection and retransmission activities. Multiple Access with Collision Avoidance [21] (MACA) and Multiple Access with Collision Avoidance Wireless [22] (MACAW) are two different MAC layer channel access protocols, aimed at avoiding or minimizing the collision rate by using a particular handshake signaling. Conversely, Power Controlled Multiple Access [23] (PCMA) is a MAC protocol that can achieve power-controlled transmission and thus collision avoidance, originally proposed for ad-hoc networks but suitable to WSNs as well.

Operating at PHY level and exploiting the frame filtering technique, it is possible to achieve a substantial reduction in energy consumption. Usually receivers perform the channel clear assessment (CCA) in order to check for incoming packets or avoid collision. The IEEE 802.15.4 standard defines three possible methods to perform the assessment:

- Energy above threshold. If the energy detected is above a fixed threshold the CCA shall report a busy medium.
- Carrier sense only. This method checks for a signal with modulation and spreading characteristics of the IEEE 802.15.4. In this case the signal may be above or below the threshold.
- Carrier sense with energy above the threshold. This is a combination of the previous methods checking both signal characteristics and energy.

Once the CCA reports a busy channel the receiver may start its RX phase to obtain the packet content. It is clear that in a dense network, where a lot of transmissions occur, there are many chances to detect a transmission. In this case most of the packets sent on the network are not intended to the receiver itself but to others receivers, generating unintentional package reception. Each package reception is an energy expensive procedure and for this reason it may be reduced at the minimum avoiding unintentional package. In [24] the authors exploit the characteristic of the Texas Instruments CC2520 RF transceiver of executing specific operations once the packet header is received. In particular, the authors modified the firmware of the RF transceiver in order to trigger an interrupt when the packet header reports as recipient an address different from its own. In this case the RF transceiver sends an interrupt to the MCU that turns off the radio module saving the energy needed to receive the packet payload. The energy consumption of CCA, result of frame dropping, unintentional and intentional package is represented in Figure 4. In subFigure 4.c, a representation is shown of the scenario where an unintentional package is received. In fact after reception the MCU does not perform any task, the opposite of what happens in scenario 4.d, where a payload processing is performed. Figure 4.b represents the case where the RF module triggers an interrupt on the MCU, saving the energy cost for the payload reception.

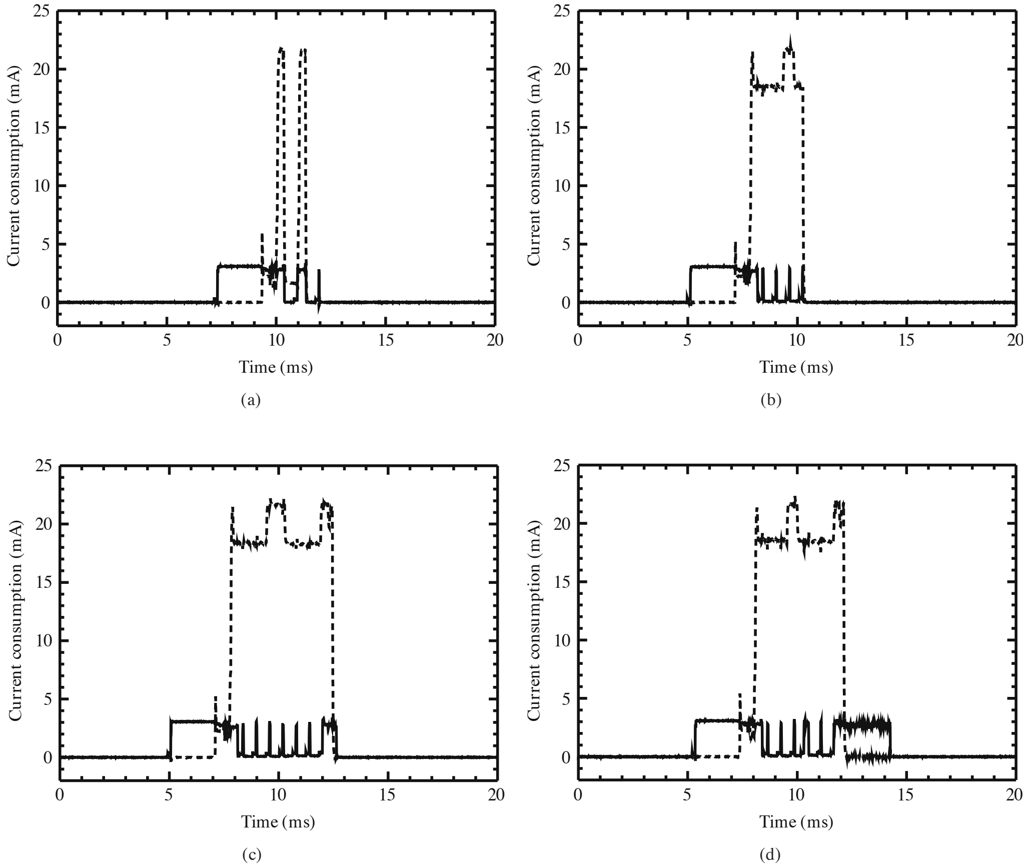


Figure 4. Power consumption of MCU (solid line) and radio transceiver (dotted line) during: (a) channel clear assessment, (b) frame filtering, (c) broadcast reception, and (d) broadcast reception and processing. [24]. Copyright © 2012 Emanuele Lattanzi and Alessandro Bogliolo, distributed under the Creative Commons Attribution License.

At routing level several mechanisms were proposed in last years to increase the lifetime of the WSN, increasing the lifetime of the nodes' batteries [3-5]. Most of the proposed techniques take into account nodes powered by batteries but the routing strategies may change taking in consideration nodes energetically sustained by their environment changing. In this scenario the routing mechanism has to be dynamically selected taking in consideration not the total energy available on the batteries but the energy available for each node at specific time (i.e. the power available). In [25] the authors demonstrate that power-constrained WSNs, as nodes powered by energy harvesting, can be represented as flow networks and that the optimization of the energetic sustainability of the workload can be cast into an instance of maxflow. Starting from this consideration, Lattanzi et al, in [26], propose a non-deterministic routing table that can be actually applied at the sensor nodes in order to achieve the maxflow theoretical optimum. In this case, since the information of the available power is accessible only to the relative node, the nodes have to cooperate to solve the maxflow problem.

4. Modeling and measurement of power consumption in WSNs

In order to ensure the expected lifetime in a WSN it is important to properly define the workflow of the nodes, evaluating and measuring their power consumption. Such evaluation may provide feedback during application design phase, consenting to improve the overall energy efficiency. The power consumption profiling of a node is also an important stage in the deployment of a WSN, since it consents to properly configure the duty-cycle and the number of transmissions as a function of the available energy. There are several methods to estimate the power consumption of a WSN node, including theoretical estimation, direct measurements, and usage simulations tools.

Theoretical estimation relies on an abstraction of the network, including the surrounding environment. However, due to the difficulty of describing the environment, realistic models are not easily realized and evaluated and even simplified models can be very complex, resulting impractical, or not accurate [27].

Direct measurements, relying on physical sensor node, offer the best accuracy on energy consumption estimation and evaluation, and are often used. Due to the complexity of the network sometime measuring the energy consumption of a whole sensor network results a very complex task. Not only measurements should be collected in different places, but WSN state and distributed power consumption measurements may require a common time reference shared by the involved nodes, so that local measurements are properly synchronized. A hybrid framework, envisions single node measurements, to be carried out with an oscilloscope or specific instrumentation under fixed conditions. Then measurement taken on a single node may be projected to the entire WSN only under some specific conditions (e.g. when the WSN nodes are homogeneous ad performs similar task). A wide measurement campaign can be carried out on limited size WSNs using specific systems [28].

Due to the variety of available platforms and environmental constraints, the design, implementation and deployment of a sensor network application are complex tasks. Thus it is often useful to simulate, at various stage of development, one or more components of the networks. Thus, accurate simulators may be a useful tool for the assessment of the WSN performance, especially given knowledge of available energy source, and the achievable duty cycle and operational life. Typical requirements are then accurate simulation of network behavior in response to specific events, accurate simulation of individual node behavior, and time awareness. Moreover, a WSN node typically includes mixed signal processing devices. As anticipated, a well known mixed signal contributor to power consumption is the radio interface, responsible for a large portion of energy consumption. Other significant contributors may be active sensors and A/D converters, whose conversion time, on a microcontroller platform, may be lower than a microsecond. In this case, not only a low level finite state machine should be modeled, but, for calibration purposes, an accurate measurement system is needed, capable of tracking state transitions lasting a few microseconds.

Specific solutions have been proposed in recent years for various wireless systems. For instance, some simulators have been developed for PAN network devices, such as Bluetooth. In [29], a Bluetooth device has been described as a finite state machine, each state being associated to link manager level activities, such as a scan/inquiry operation. Then, average power consumption was measured for each of the identified state transitions, each lasting a few milliseconds, using Digital Multi Meters (DMM). As a result, the average power consumption of a Bluetooth device executing a given application could be predicted with good accuracy [29]. While effective, such approach features a large time granularity that may be suboptimal for WSN applications. In fact, WSN nodes are often arranged in a peer to peer or mesh configuration, where several asynchronous and short events may occur, and featuring various low-power/sleep modes. Moreover, a deep optimization of power consumption may require a simulation tool to profile the energy cost of the internal work of each node. This requires to model events with time constants that may be lower than a microsecond [30].

Thus, other WSN simulators have been recently developed, focused on the simulation of the protocol and MAC level, on processor profiling, on attempting to combine both features [31]. In network focused simulation frameworks, sensor nodes are generally represented using a layered architecture, where each layer is responsible to model specific hardware or software aspect of the node. Moreover, in order to study the energy consumption profile of the node, accurate timing information is needed. Thus, such simulators alone, being oriented to model the network activity and the information flow, lead to a coarse representation of the node states, and are not suitable for accurate energy consumption estimation. Another class of simulators emulates the platform executing the same code of the node. Using this technique it is possible to obtain a fine-grained timing, permitting the simulation of interrupts and low-level peripheral interaction. Such simulators are usually called instruction-level simulators. Due to the strict hardware dependence each simulator is usually capable to emulate only a few platforms, relying on configuration files that describe the peculiar characteristics of a given platform. It should be noted that, in order to obtain an accurate simulation of power consumption, also platform components embedded in the WSN node with the CPU may be significant contributors to power consumption, and should be properly kept into account. An example is provided in section 5, where a case study is discussed.

Finally, simulation tools should be coupled to proper measurement techniques, keeping into account potential and limitation of the available instrumentation. For instance, in [29] accurate measurements have been carried out using DMMs, and measurement uncertainty has been modeled by describing the effect of measuring phenomena of comparable duration with the DMM integration time. More generally, the requirements of a measurement system should include accuracy, and the capability of capturing and measuring phenomena with short duration and a potentially low repetition frequency.

In the next subsections different methods to evaluate by experiment or by simulation the energy consumption of a WSN node are presented.

4.1. Measurement setup for current consumption

Since embedded systems usually operate at constant supply voltage, power consumption measurements can be carried out indirectly, by measuring and monitoring the absorbed current. To this aim, various techniques are available, described in the following.

A very common solution is the series insertion of a small resistance R_1 (≤10Ω) between the power supply and the Device Under Test (DUT), as shown in Figure 5. Then, by measuring the voltage drop ΔV=V_2-V_1 across the resistor, current I can be measured indirectly, using Ohm's law, as

$$\hat{I} = \frac{\Delta V}{R_1} = \frac{V_2 - V_1}{R_1} \tag{1}$$

and the absorbed power can be estimated as $\hat{P} = \hat{I} \cdot V_0$.

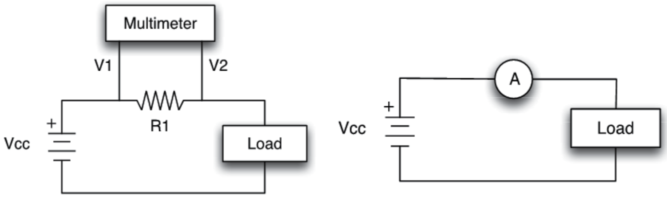


Figure 5. Measurement setup with a shunt resistor or amperemeter

Such measurements, performed during normal operation of the platform, allow monitoring the power consumption when different components of the board are active, and its time dependence. A key role in the reliability of such measurement is played by the voltage measurement system accuracy, the tolerance on the shunt resistor value, the stability of the supply voltage, and the measurement rate, that should be compatible with the analyzed phenomena. Notice that, with such a system, the voltage drop across the resistor reduces the supply voltage powering the DUT, introducing a type B contribution to the measurement uncertainty for large values of R_1 [32]. On the other hand, low values of R_1 lead to reduced values of ΔV, resulting in lower measurement sensitivity. By assuming that the aforementioned offset effect is negligible, and that uncertainties on R_1, V_0, and ΔV are uncorrelated, using the law of propagation of uncertainties the measurement uncertainty $u(I)$ on current I and the measurement uncertainty $u(P)$ on power consumption P are respectively given by [32]

$$u(I) \cong \sqrt{u(\Delta V)^2 \frac{1}{R_1^2} + u(R_1)^2 \left(\frac{\Delta V}{R_1^2} \right)^2} \tag{2}$$

and

$$u(P) \cong \sqrt{u(\Delta V)^2 \left(\frac{V_0}{R_1} \right)^2 + u(R_1)^2 \left(\frac{V_0 \Delta V}{R_1^2} \right)^2 + u(V_0)^2 \left(\frac{\Delta V}{R_1} \right)^2} \tag{3}$$

where $u(\Delta V)$, $u(V_0)$, and $u(R_1)$ are uncertainties on ΔV, V_0, and R_1 respectively.

A similar approach to the resistor method is the direct insertion of an amperemeter, capable of measuring currents ranging from the microampere, the typical absorption of a microcontroller settled in sleep mode, to a few tens of milliampere, corresponding to a full workload (data collection/processing, RF transmissions). Uncertainty contributions may be evaluated using an approach similar to that associated to the resistor method. Notice that, depending on the amperemeter architecture, bandwidth limitations may lead to averaging of the measured current, leading to a loss of information [30].

In order to improve the measurement accuracy, alternative approaches have been suggested. For instance, in [33] a method has been proposed, based on inserting a switched pair of capacitors between the power supply and an ARM7TDMI processor, as shown in Figure 6. By alternatively switching the capacitors C_{S1} and C_{S2} with the microcontroller clock, the processor can be powered by the capacitors. By also keeping into account the effect of the on-chip capacitance, the energy consumption can thus be estimated by measuring over time the voltage drops across both capacitors, and by recalling that the energy stored in a capacitance C with a

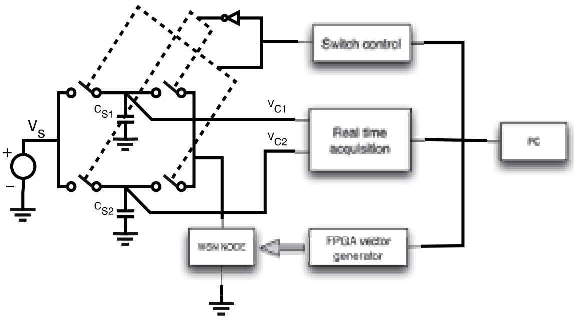


Figure 6. Measurement setup with switch capacitor

voltage drop *V* is given by $CV^2/2$. This method removes the offset uncertainty introduced by the resistor method.

Another method to reduce measurement uncertainty has been proposed in [30]. Here, a current mirror has been designed as shown in Figure 7, whose symmetric topology replicates the current absorbed by the microcontroller. Such replica is then measured, without perturbing the microcontroller power absorption. In this case accuracy is limited by tolerances of the current mirror components, which should be carefully matched in order to guarantee an accurate replication of the current absorbed by the DUT.

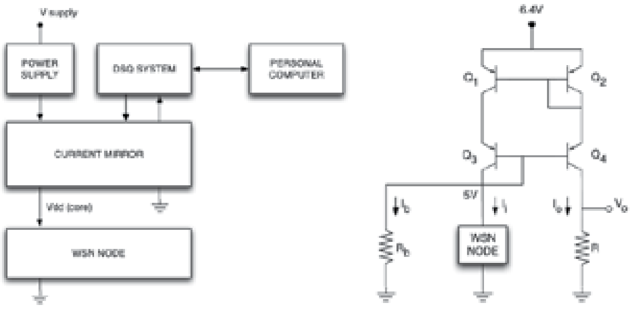


Figure 7. Diagram and circuit of measurement setup with current mirror

It should also be observed that the proposed approaches are suitable for node level measurements, but may be unpractical in large WSNs, especially if deployed in large of harsh environments. In this case specific solutions are used. Hergenröder et al. in [28] presented a distributed energy measurement system called Sensor Node Management Device (SNMD), used in conjunction with the SANDbed testbed [34]. SNMD is a measurement system suitable for different sensor node (e.g. MicaZ, IRIS, SunSPOT), providing energy measurement on individual nodes. The current measurement is based on shunt resistor (i.e. 1 Ohm) approach as described before. The resulting system has a current selectable range up to 500mA on 0-10 Voltage range with 16bits of resolution and a sampling rate up to 500kHz (20kHz without buffering). A key aspect in distributed measurement, using a SNMD device on each WSN node, is the synchrony between the measurements on different nodes. For this reason on SNMD the time is synchronized using the Network Time Protocol (NTP) that provides an accuracy of 10ms. However such accuracy can be too coarse for specific applications, especially in relation to the high sampling frequency of the node. For this reason the authors in [28] proposed the use of offline algorithms to synch and analyze measurements performed on different nodes.

An effective measure of current consumption of a WSN node during operation is presented in Figure 8.

4.2. Modern energy consumption simulation software

As mentioned before, the simulation software depends on the considered node architecture. In this section we present two different instruction level simulation software, developed

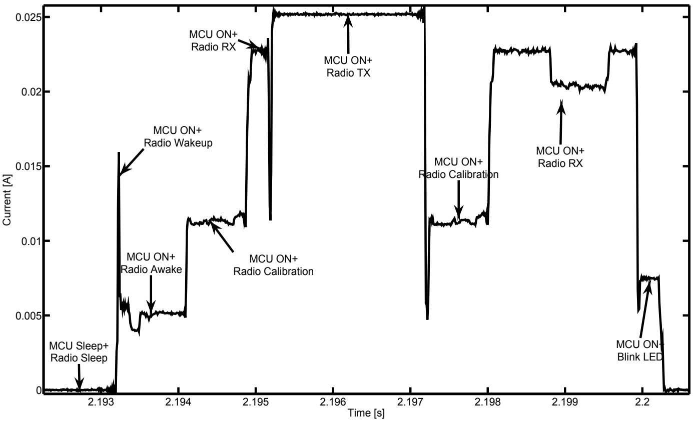


Figure 8. Current consumption measurements for a WSN node during transmission with acknowledgement

respectively for the AVR microcontroller, produced by Atmel, and for the MSP430 microcontroller, produced by Texas Instruments.

The first considered simulator, known as Avrora, is a set of simulation and analysis tools developed the UCLA Compilers Group [35]. In particular, the simulator can handle up to a few thousands nodes, by taking advantage of the processing power of modern computers. Avrora is not only a simulator to test program execution on the node but it also allows online monitoring of the code execution on the WSN, profiling utilities to study the program's behavior, source level debugging, a control flow graph, providing a graphical representation of program's instructions, and an energy analysis tool, capable of analyzing the energy consumption of a specific application.

The simulator has been enhanced by Haas et al., after evaluating the performance of the analysis tools of Avrora, comparing the simulation results with experimental measurements carried out with the SANDbed platform [36]. The test application, run over TinyOS and involving four nodes with fixed routing path. Using the collected data Haas et al. developed and released an enhanced version of Avrora, called Avrora+, improving the calibration of energy model, modeling transition state cost, and taking into account the effect of manufacturing tolerance on the energy consumption.

The experimental verifications showed that the Avrora+ is very accurate, reducing the difference between measurements consumption measurements and simulation results to less than 5%.

The Worldsens simulation Framework is another WSNs simulator that support MSP430 based node [37]. This open source platform, released under the CeCILL and GNU GPL license agreements, includes three simulation tools, often used in conjunction:

- WSim: this is the platform simulator. It performs a full instruction simulation of the node, driven by the microprocessor internal clock.

- WSNet: an event driven wireless network simulator that can be used in conjunction with WSim to simulate a whole sensor network with high accuracy.
- eSimu: a software module that implements platform specific energy consumption models, and provides an estimation of the current absorbed by a node [38].

For estimating the power consumption, WSim and eSimu are usually being jointly used, interacting as shown in Figure 9. In particular, the WSim tool, compiled with eSimu support, receives the binary file that would be executed by a real microprocessor for a given application and provides a trace file, describing state transitions of the node and its peripherals. Notice that, when modeling a radio transmission, WSNet is used as well, in conjunction with WSim. Using the trace file and a calibration file reporting the current absorbed by the node in its various states, the overall current consumption of a node that executes a given task can be estimated and profiled against the execution time. Since the node is powered by a constant voltage source, the power consumption can easily be derived from the current absorption.For an exhaustive overview of existing tools for simulations, modeling and measurements of WSNs refer to [39].

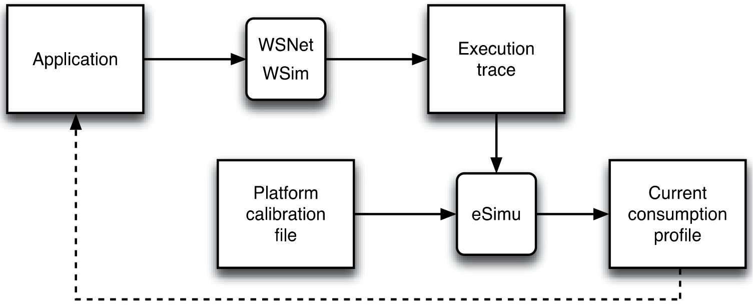


Figure 9. Wsim, WSNet and eSimu simulation process

5. Case study

In this section we present experimental measurements along with simulation results regarding one of the platforms mentioned in section 2: the ez430-RF2500 development kit from Texas Instruments. The considerations discussed remain valid for similar nodes and may be applied to different architectures or cases.

5.1. Platform description

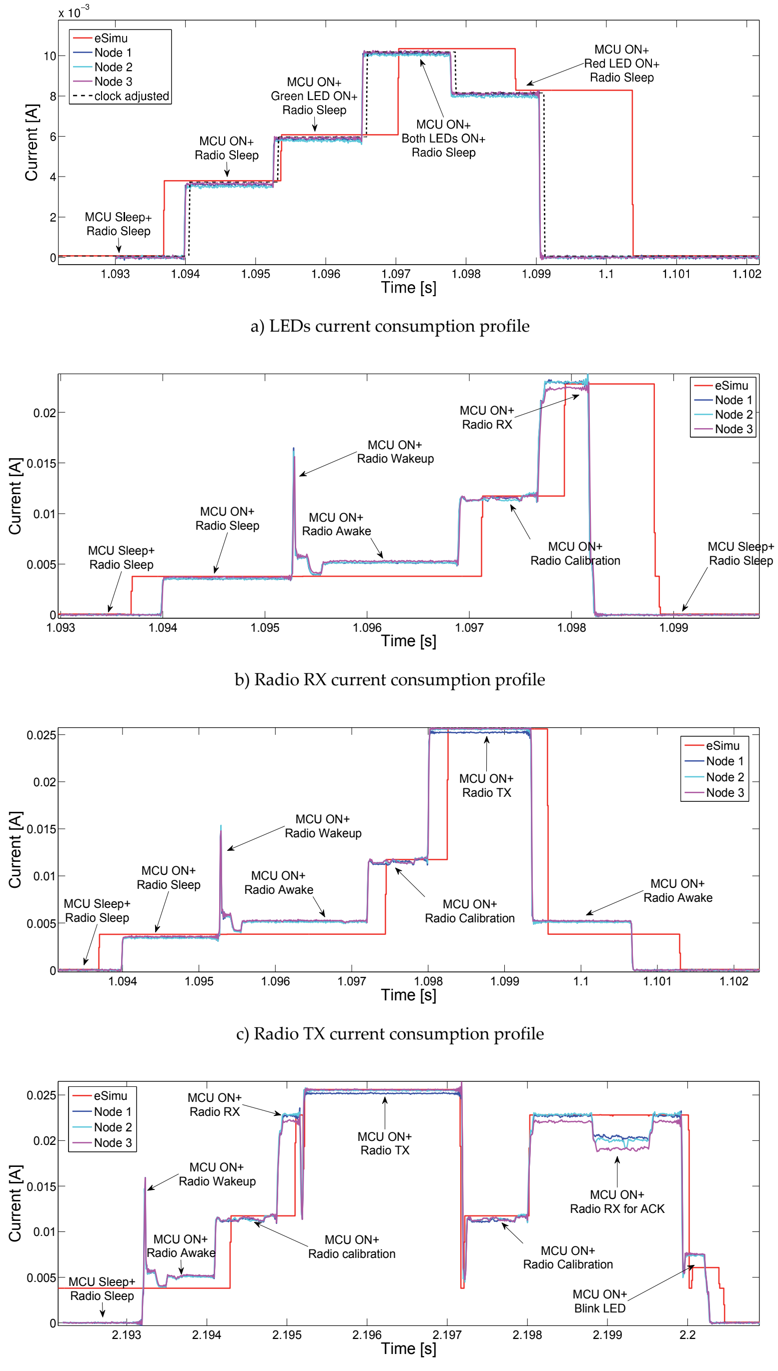
The ez430-RF2500 is a development tool for WSNs produced by Texas Instruments, which includes a MSP430 low-power microprocessor and a CC2500 radio module. External sensors can be connected to the board through a Serial Peripheral Interface (SPI) port, an I2C interface, or a 10 bit Analog to Digital Converter (ADC) with a sampling rate of 200ksps. The standard operational voltage of the platform is 3.6V, by factory the ez430-RF2500 board is powered by an external battery pack (2 AAA batteries). The platform core is an ultra low power micro-controller of the MSP430 family with 32kB of ROM, 1kB of RAM and a clock up to 16MHz [40]. Communication between nodes has been implemented using the CC2500 radio module. The CC2500 radio module is a 2.4 GHz transceiver with a low power hardware wake up function. The transceiver embeds a highly configurable baseband modem, which supports various modulation formats and a configurable data rate up to 500kBd. The Radio Frequency (RF) module does not support Offset Quadrature Phase-Shift Keying (O-QPSK) and for this reason is compliant to the 802.15.4 IEEE standard. Thus it is not possible to implement the ZigBee specification in this specific platform. Texas Instruments developed an alternative network protocol, similar to ZigBee, called SimpliciTI. The SimpliciTI network protocol is a low-power radio-frequency protocol targeting simple, small RF networks whit less than 100 nodes. The SimpliciTI network protocol is designed for easy implementation with minimal micro-controller resource requirements, and supports End Devices in a peer-to-peer network topology, also permitting the usage both of an Access Point to store and forward messages to a LAN, and of Range Extenders, that may extend the range of the network up to four hops [41].

5.2. Current consumption experiment and simulations

In Figure 10 several measurements along with simulation are presented in order to highlight the energy consumption of the examined board in different scenarios. The real current consumption profile has been measured for three different ez430-RF2500 nodes. The measurements have been performed using the shunt resistor technique, as presented in the section 4.1, at V_0=3.3 V using as voltage measurement system a National Instruments USB-4432 Data Acquisition System (DAS). The presented setup guaranties a quantization step of 4.77 μV, given the 24 bits resolution of the DAS and its voltage range (±40 V), resulting in 0.477 μA current resolution with a 10 Ohm resistance. The simulated data has been produced using the Worldsens simulation Framework, presented in section 4.2, applying some modification in order to better model the ez430-RF2500 platform. Such modifications consist in:

- A better model of the platform including as actuator the two LEDs present on the board (each one with different typical current consumption).
- An optimized energy analysis function capable of analyzing trace execution file with high resolution for long execution periods.
- A scale factor to match correctly the execution time; the simulated clock doesn't match the real system clock.
- An energy evaluation of the radio from sleep to wake-up transition.

In Figure 10.a the current consumption of the platform with the LEDs in use is represented. As evident there is a good agreement between the simulated data and the experimental results. In this scenario the platform starts from a condition where booths LEDs are turned on and the Micro-Controller Unit (MCU) in low power mode (LPM3). Then the boards wake up and turn on the LEDs, finally going to sleep again. Notice that the dashed black line represents the simulated current consumption applying the suggested clock adjustment and the solid red curve without the clock adjustment.



a) LEDs current consumption profile

b) Radio RX current consumption profile

c) Radio TX current consumption profile

d) TX with ACK request current consumption profile

Figure 10. Current consumption experimental measurements and simulation results

The current consumption contribution of the RF module is evident in Figure 10.b and 10.c. In those scenarios the platform starts from a LPM3 and after the wake up of the microcontroller it turns on the radio module to perform respectively a reception (RX) and a transmission (TX). It is evident that in such condition the RF module consumes the greatest part of the energy. Even in this case there is a good agreement between experiment and simulation, however there is evident a mismatch between simulations and measurements during the wake up of the radio: The spike measured is due to the current absorption by the crystal oscillator and the WSim software does not take into account such analog phenomena.

The last considered scenario (represented in Figure 10.d) is a bidirectional communication between two nodes using the SimpliciTI network protocol. The node starts in a sleep state, and then it wakes up, turns on the radio, and sends a 40 *char* message to the Access Point (AP) using the SimpliciTI function *SMPL_SendOpt*. The message is sent requesting an acknowledgment in order to confirm a successful transmission. To accomplish the transmission with the acknowledged reception, the RF module switches from the IDLE state to RX state to check if the channel is free, and then starts the transmission. After its completion the node reverts to the RX state, repeating the calibration step and wait for the acknowledgment. Once received the LED blinks to notify the reception and the MCU returns to sleep mode.

6. Conclusions

A review on power consumption measurements in WSN networks has been presented, highlighting the main WSN features, the node architecture, and the network operation. Measurement and simulation techniques adopted to assess the power consumption of a WSN node have been discussed, showing the most significant approaches, the underlying tradeoffs of each methodology, and discussing the achievable accuracy. A case study has been introduced, presenting a characterization procedure and developing improvements for an existing WSN simulator.

Author details

Antonio Moschitta[1] and Igor Neri[2*]

*Address all correspondence to: igor.neri@nipslab.org

1 Department of Electronic and Information Engineering, University of Perugia, Perugia, Italy

2 NiPS Laboratory, Department of Physics, University of Perugia, Perugia, Italy

References

[1] IEEE 802.15TG4 Web site, Available: http://www.ieee802.org/15/pub/TG4.html – Consulted December 2012

[2] "IEEE Standard for a Smart Transducer Interface for Sensors and Actuators". IEEE Standards Association. doi:10.1109/IEEESTD.2007.4346346. 2009 – Consulted December 2012

[3] Camilo, Tiago, et al. "An energy-efficient ant-based routing algorithm for wireless sensor networks." Ant Colony Optimization and Swarm Intelligence(2006): 49-59.

[4] Chiti, Francesco, et al. "Energy efficient routing algorithms for application to agro-food wireless sensor networks." Communications, 2005. ICC 2005. 2005 IEEE International Conference on. Vol. 5. IEEE, 2005.

[5] Abbasi, Ameer Ahmed, and Mohamed Younis. "A survey on clustering algorithms for wireless sensor networks." Computer communications 30.14 (2007): 2826-2841.

[6] Tayal, Animesh R., N. V. Choudhary, and Madhuri A. Tayal. "Simulation of sensor nodes for energy conception in wireless sensor networks using finite automata." Proceedings of the International Conference on Advances in Computing, Communication and Control. ACM, 2009.

[7] Barboni, Leonardo, and Maurizio Valle. "Experimental analysis of wireless sensor nodes current consumption." Sensor Technologies and Applications, 2008. SENSORCOMM'08. Second International Conference on. IEEE, 2008.

[8] Tennina, Stefano, et al. "Locating zigbee® nodes using the ti® s cc2431 location engine: a testbed platform and new solutions for positioning estimation of wsns in dynamic indoor environments." Proceedings of the first ACM international workshop on Mobile entity localization and tracking in GPS-less environments. ACM, 2008.

[9] Aamodt, K. "CC2431 location engine–Application note AN042." Chipcon products for Texas Instruments (2008): 20.

[10] Orfei, F. et al. "HYBRID AUTONOMOUS TRANSCEIVERS", EDERC2012: 5th European DSP Education and Research Conference, 13/09/2012, Amsterdam, (2012)

[11] Seah, Winston KG, Zhi Ang Eu, and Hwee-Pink Tan. "Wireless sensor networks powered by ambient energy harvesting (WSN-HEAP)-Survey and challenges." Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, 2009. Wireless VITAE 2009. 1st International Conference on. IEEE, 2009.

[12] Vullers, Ruud JM, et al. "Energy harvesting for autonomous wireless sensor networks." Solid-State Circuits Magazine, IEEE 2.2 (2010): 29-38.

[13] Potdar, Vidyasagar, Atif Sharif, and Elizabeth Chang. "Wireless sensor networks: A survey." Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on. IEEE, 2009.

[14] Farooq, Muhammad Omer, and Thomas Kunz. "Operating systems for wireless sensor networks: a survey." Sensors 11.6 (2011): 5900-5930.

[15] G. Gupta and M. Younis, "Load-balanced clustering of wireless sensor networks", in Proceedings of 2003 IEEE International Conference on Communications (ICC'03), Anchorage, AK, May 2003, pp. 1848-1852.

[16] R. Rajagopalan and P. Varshney, "Data-aggregation techniques in sensor networks: A survey", IEEE Communications and Surveys and Tutorials, vol. 8, no. 4, 4th Quarter 2006, pp. 48-63.

[17] A. A. Abbasi and M. Younis, "A survey on clustering algorithms for wireless sensor networks", Computer Communications, vol. 30, nos. 14-15, Oct. 2007, pp. 2826-2841.

[18] C. Srisathapornphat and C. C. Shen, "Coordinated power conservation for ad-hoc networks", in Proceedings of 2002 IEEE International Conference on Communications (ICC'02), vol.5, New York, Apr.-May 2002, pp. 3330-3335.

[19] C. Imrich, H. Kim, and S. Ha, "Dynamic voltage scheduling technique for low-power multimedia applications using buffers", in Proceedings of the International Symposium on Low Power Electronics and Design, Huntington Beach, CA, June-July 2001, pp. 34-39.

[20] A. Sinha and A. P. Chandrakasan, "Dynamic power management in wireless sensor networks", IEEE Design and Test of Computers, vol. 18, no. 2, Mar.-Apr. 2001, pp. 62-74.

[21] P. Karn, "MACA - A new channel access method for packet radio", in Proceedings of the ARRL CRRL Amateur Radio 9th Computer Networking Conference, Redondo Beach, CA, Apr. 1990, pp. 134-140.

[22] V. Bharghavan, A. Demers, S. Shenkar, and L. Zhang, "MACAW: A media access protocol for wireless LANs", in Proceedings of ACM SIGCOMM'94, London, UK, Sept. 1994, pp. 212-225.

[23] J. Monks, V. Bharghavan, and W. M. Hwu, "A power controlled multiple access protocol for wireless packet networks", in Proceedings of IEEE INFOCOM'01, vol. 1, Anchorage, AK, Apr. 2001, pp. 219-228.

[24] Lattanzi, Emanuele, and Alessandro Bogliolo. "VirtualSense: A Java-Based Open Platform for Ultra-Low-Power Wireless Sensor Nodes." International Journal of Distributed Sensor Networks 2012 (2012).

[25] A. Bogliolo, E. Lattanzi, and A. Acquaviva. "Energetic sustainability of environmentally powered wireless sensor networks". In PE-WASUN '06: Proceedings of the 3rd

ACM international workshop on Performance evaluation of wireless ad hoc, sensor and ubiquitous networks, pages 149– 152, 2006.

[26] E. Lattanzi, E. Regini, A. Acquaviva, and A. Bogliolo, "Energetic Sustainability of Routing Algorithms for Energy-Harvesting Wireless Sensor Networks", Elsevier Computer Communications, Special Issue on Network Coverage and Routing Schemes for Wireless Sensor Networks, Vol. 30, No. 14-15, pp. 2976-2986, 2007.

[27] Abuarqoub, Abdelrahman, et al. "Simulation Issues in Wireless Sensor Networks: A Survey." SENSORCOMM 2012, The Sixth International Conference on Sensor Technologies and Applications. 2012.

[28] Hergenröder, Anton, Joachim Wilke, and Detlev Meier. "Distributed energy measurements in WSN Testbeds with a sensor node management device (SNMD)." Architecture of Computing Systems (ARCS), 2010 23rd International Conference on. VDE, 2010.

[29] D. Macii, D. Petri, "An Effective Power Consumption Measurement Procedure for Bluetooth Wireless Modules," IEEE Transactions on Instrumentation and Measurements, Vol. 56, no. 4, August 2007.

[30] T. Laopoulos, P. Neofotistos, C. A. Kosmatopoulos, and S. Nikolaidis," Measurement of Current Variations for the Estimation of Software-Related Power Consumption," IEEE Transactions on Instrumentation and Measurements, Vol. 52, no. 4, August 2003.

[31] Egea-Lopez, E., et al. "Simulation tools for wireless sensor networks."Proceedings of the International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'05). 2005.

[32] Bich, Walter, Maurice G. Cox, and Peter M. Harris. "Evolution of the'Guide to the Expression of Uncertainty in Measurement'." Metrologia 43.4 (2006): S161.

[33] Chang, Naehyuck, Kwanho Kim, and Hyung Gyu Lee. "Cycle-accurate energy measurement and characterization with a case study of the ARM7TDMI [microprocessors]." Very Large Scale Integration (VLSI) Systems, IEEE Transactions on 10.2 (2002): 146-154.

[34] Hergenröder, Anton, Jens Horneber, and Joachim Wilke. "SANDbed: A WSAN Testbed for Network Management and Energy Monitoring." Hamburg, Germany, Aug 8 (2009).

[35] Titzer, Ben L., Daniel K. Lee, and Jens Palsberg. "Avrora: Scalable sensor network simulation with precise timing." Information Processing in Sensor Networks, 2005. IPSN 2005. Fourth International Symposium on. IEEE, 2005.

[36] Haas, Christian, Joachim Wilke, and Viktor Stöhr. "Realistic simulation of energy consumption in wireless sensor networks." Wireless Sensor Networks(2012): 82-97.

[37] G. Chelius, A. Fraboulet, E. Fleury, ”Worldsens: development and prototyping tools for application specific wireless sensors networks”, in International Conference on Information Processing in Sensor Networks (IPSN) (Boston, USA), ACM, April 2007.

[38] N. Fournel, A. Fraboulet, and P. Feautrier, “Embedded Software Energy Characterization: using non-intrusive measures to annotate application source code”, Journal of Embedded Computing, Volume 3 (3), IOS Press, July 2009

[39] A. Dwivedi and O. Vyas, “An Exploratory Study of Experimental Tools for Wireless Sensor Networks”, Wireless Sensor Network, Vol. 3 No. 7, 2011, pp. 215-240. doi: 10.4236/wsn.2011.37025.

[40] Texas Instruments, eZ430-RF2500 Development Tool User Guide, Literature Number: SLAU227E September 2007 – Revised April 2009

[41] Texas Instruments, “SimpliciTI Compliant Protocol Stack.” 2009207210. http://www.ti.com/tool/simpliciti – Consulted December 2012.

Autonomous Sensors: Existing and Prospective Applications

Francesco Orfei and Giulia Orecchini

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/57348

1. Introduction

The progressive diffusion of wireless sensor networks [1] implies an increase in batteries utilization and a consequent concern about their proper disposal at the end of their lifetime. Together with the reduction of devices dimensions, the decrease of the available space for electronics and batteries generally implies a shorter autonomy of operation compared to previous solutions.

Moreover the battery's self-discharge can be an issue when it has to be stored in the device for a long time, as it usually happens in the case of sensor applications. Thus an alternative power supply technology that provides a viable solution towards the improvement in terms of operational efficiency is becoming one of the main interesting topics concerning the realization of wireless sensor networks.

There are several ways to have the sensor always ready to operate, considering that rechargeable batteries can last longer if recharged by an alternative power source. Standard rechargeable batteries can be paralleled to an alternative power source like a photovoltaic (PV) cell or a piezoelectric generator through a battery charger. In some case one or more power sources can be used at the same time.

Given these ideas, it would be possible to have an autonomous wireless sensor powered only by the energy coming from the environment: no batteries on board would be required. This approach is now becoming popular as “energy harvesting technology” [2]. In this chapter we focus on an innovative solution that combines a PV cell and a piezoelectric or an electromagnetic vibration energy harvester. These generators use different sources of energy and they can be complementary for example in situations where one of the two may be temporarily unavailable.

It can be noticed that the device overall cost does not depend only on the development and production cost, but for a battery powered device, the costs of maintenance have to be considered as well. These costs are clearly related to the device performances; the more it consumes the more often battery replacement and disposal is required. As a consequence, at least three more costs have to be considered: the cost for battery purchase, the cost for battery replacement and the cost for its disposal.

It is evident that a system that does not require any battery is a very good candidate for the development of future wireless sensors networks.

2. A typical autonomous sensor

Generally one or more power sources can be used at the same time; for example in the block diagram shown in Figure 1 a vibration and a solar energy harvester is represented. It converts some of the environmental energy present in the working area of the sensor to electricity that can be used to power the sensor itself. This means that the sensor can virtually work forever.

The generated current from the harvester can be direct, alternate or sometimes can have a random behavior, while the internal sensor circuitry expects to have a constant value for the incoming current; therefore a power conditioning circuitry is needed to rectify and to regulate the voltage coming from the energy sources.

After the power conditioning, a voltage supervisor is required to properly turn ON or OFF each electronic part of the sensor. Without this component an electronic device, like the microcontroller, can be turned ON and OFF continuously without even starting to work properly.

A microcontroller is generally used as the processing unit of the sensor. It is devoted to the acquisition, the processing and the transmission of the information coming from the environment and collected by the sensor, through a radio frequency transceiver. One or more environment variables can be acquired by the microcontroller. Some examples are depicted in Figure 1: voltages, temperatures, light levels and many more.

The microcontroller can work with both analog and digital signals allowing the system to collect a large variety of information. Once the data have been acquired they are serialized and then sent to the radio frequency transceiver for the transmission to the remote receiver. At the end of the transmission the processor can turn OFF all the devices and itself too, to save energy. After a pre-established time, the microcontroller turns ON and the acquisition and the transmission of the variables values starts again.

3. Power supply

As mentioned before, each electronic device requires an electrical power source to work. Generally batteries are used to provide enough power to the sensor. The size of the battery

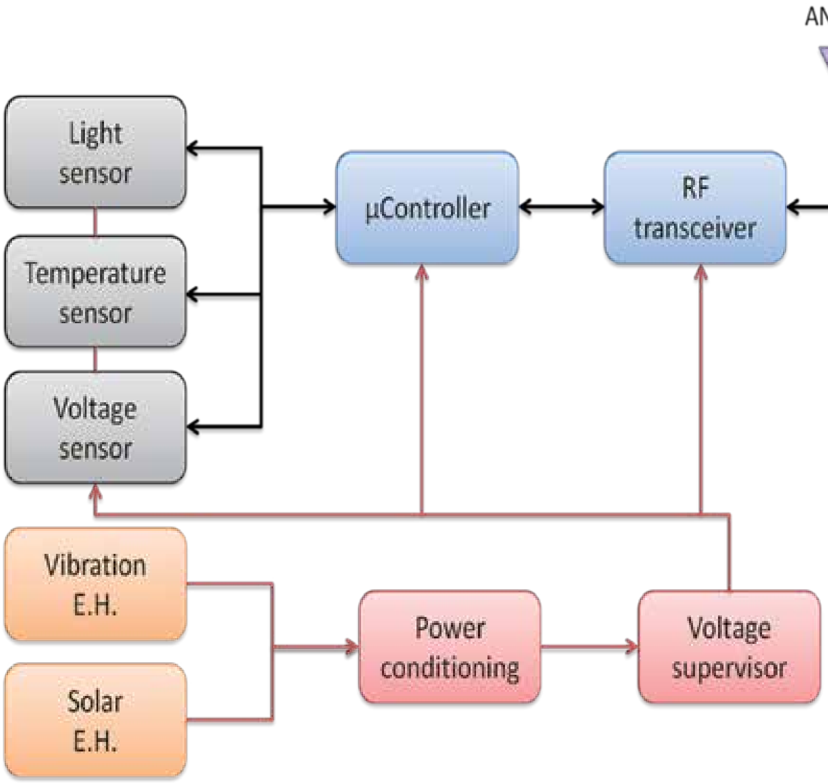


Figure 1. Block diagram of an autonomous wireless sensor

depends on the expected life of the system and on the power required to operate. An autonomous system, instead, works without batteries or any other traditional power source. This means that the required power have to be provided by itself. Energy cannot be produced, but it can only be transformed, or transferred, from a state to another state.

The task of an energy harvester is to transform the energy available from the environment to electricity. Some different sources of energy are suitable for this purpose: vibrations, thermal gradients, light and electromagnetic fields. The choice to use a source with respect to another one is related to the amount of energy available. Generally the energy extracted is not constant: sometimes it fluctuates daily, as the solar light, some other times it varies more rapidly, like the vibrations of a wheel on the road.

Since an electronic device requires a constant power supply, a storage device is needed so that the energy will be constantly available. A storage device can be realized by a low self-discharge high capacitance capacitor or a thin film rechargeable solid-state battery.

Moreover it is very important to reduce the energy wasted during the conversion, by increasing its efficiency, to be able to store as much energy as possible. In this way, given a fixed power requirement, the same amount of energy would be given by a smaller harvester resulting in a relevant size and cost reduction of the overall system.

3.1. Energy harvester

Several types of energy harvester can be realized and used to obtain the electrical energy required by a sensor to be energetically autonomous. There's no way to say which one is the best, what is possible to say is which one is the best for a particular application. The choice is not easy because a lot of parameters have to be evaluated, but first it is possible to choose among these few types:

- Vibration energy harvester
- Solar light energy harvester
- Thermal gradients energy harvester
- Electromagnetic energy harvester

Each harvester can be realized with different technologies, so the number of possible ways to make it, provides a good freedom of choice.

It is also possible to combine more energy harvesters to obtain a higher probability of having a continuous power supply. It can happen that a source of energy is temporarily unavailable but another one can still be present (for example vibration and solar energy).

Vibration energy harvesters can be realized using different technologies and materials.

- Piezoelectric energy harvesters: a piezoelectric material is stressed by vibrations to produce an electrical current proportional to their intensity. Different shapes and sizes for the piezoelectric material can be used, from a simple cantilever (Figure 2) to a buckled beam to something more complicated like a fractal.

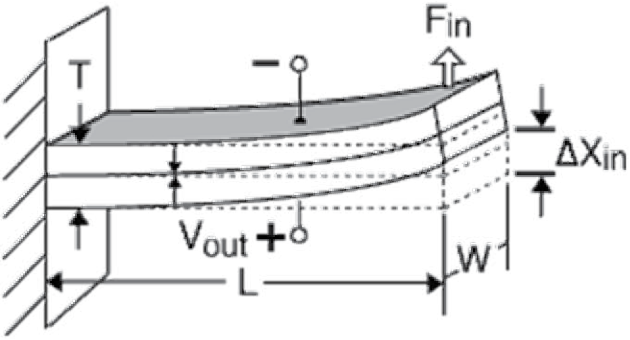


Figure 2. A bilayer piezoelectric cantilever

- Magnetostrictive energy harvesters: a magnetostrictive material is deformed by vibrations to produce a variable magnetic field that, when coupled with an inductor (Figure 3), can produce an electrical current proportional to the intensity of them.

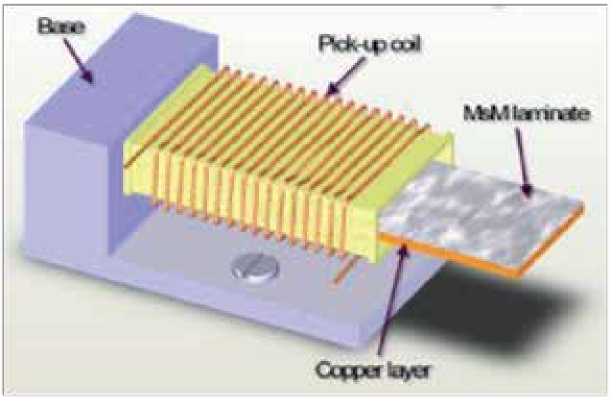


Figure 3. Magnetostrictive vibration energy harvester

- Induction based energy harvesters: vibrations generate a relative movement of a magnet and of an inductor (Figure 4). Thanks to the law of Faraday-Neumann an electrical current is generated across the inductor.

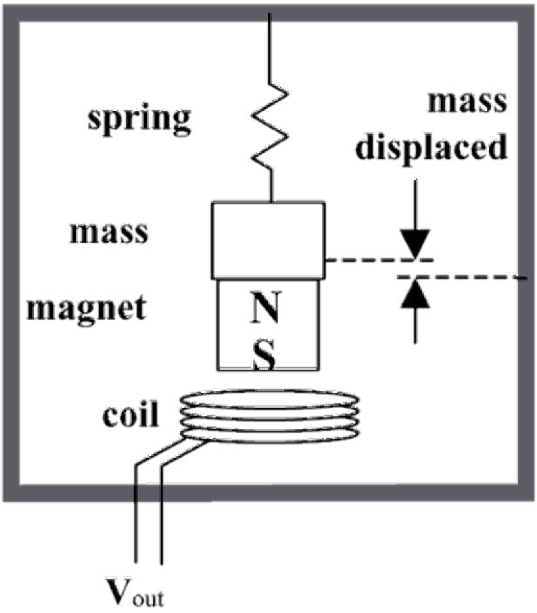
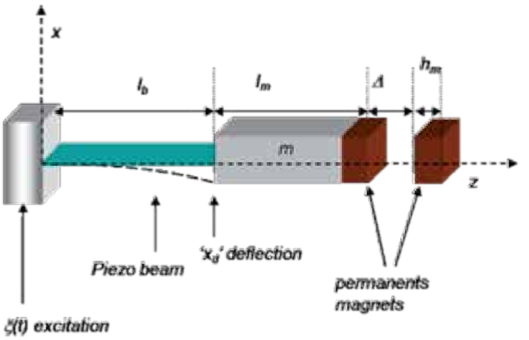


Figure 4. Schematic of an induction based energy harvester

When a vibration harvester has to be designed, it is also important to consider its dynamics. In several years of research, it has been demonstrated that the best results come with a non-linear dynamic system. The energy harvester based on bi-stable non-linear dynamics is one of the simplest (Figure 5).



$$\ddot{x} = -\frac{dU(x)}{dx} - \gamma\,\dot{x} - K_v V - \sigma\,\xi(t)$$

$$\dot{V} = -K_c\,\dot{x} - \frac{1}{\tau_p} V$$

Figure 5. Bi-stable non-linear piezoelectric energy harvester

Considering a cantilever, a non-linear bi-stable dynamics can be obtained using two magnets [3]: this harvester is shown in Figure 5. *U(x)* is the potential energy of the cantilever, *V* is the generated voltage, γ is the damping constant, K_c and K_v are coupling constants and τ_p is the time constant of the piezoelectric. Δ is the distance between the magnets and it controls the height of the barrier of the potential: refer to Figure 6.

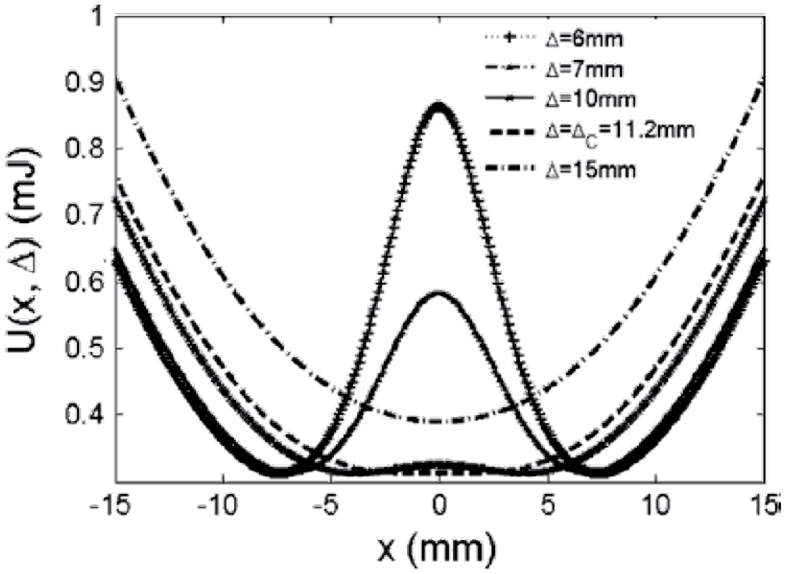


Figure 6. Potential function of a bi-stable non-linear pendulum

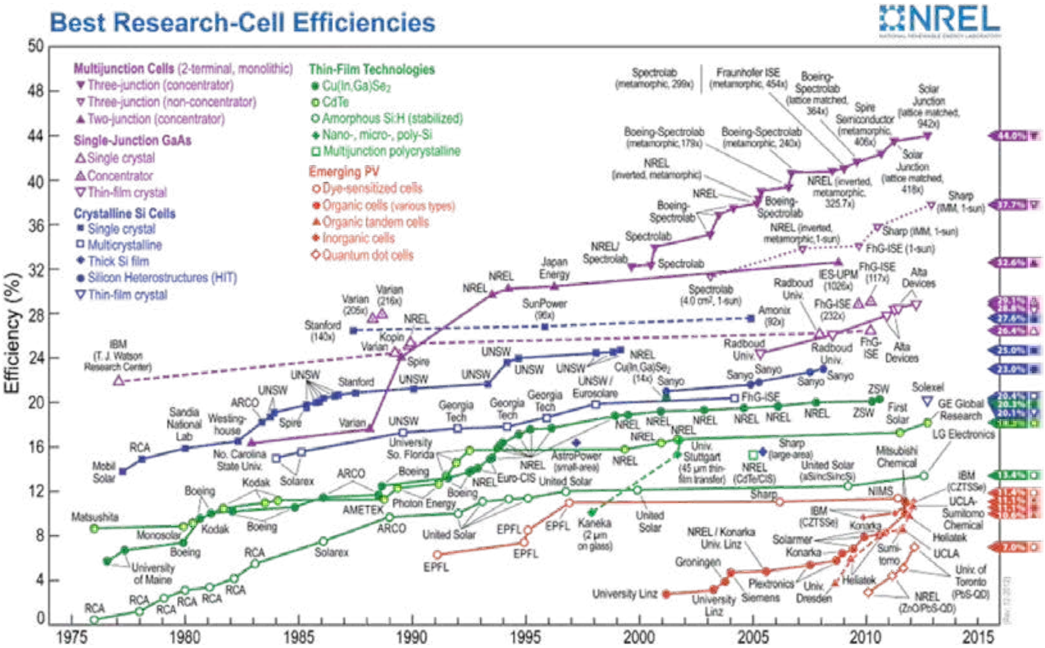


Figure 7. Timeline of solar cell energy conversion efficiencies (from National Renewable Energy Laboratory - USA)

The current produced by a vibration energy harvester is an alternate current, so it must be rectified to be used by a sensor. Solar energy harvesters, instead, use the photovoltaic effect to produce a direct current; this does not need to be rectified. Several different technologies are used and studied to produce new solar cells.

Given that the radiation power arriving on the surface of the earth is approximately 1kW/m^2, the only way to increase the power converted from light is to increase the efficiency of conversion. In Figure 7 the conversion efficiency with respect to the material used for the fabrication from 1976 to 2012 is reported.

Another source of energy is represented by thermal gradients and it is used by the so called thermoelectric generator, or TEG. Usually the Seebeck effect is used to produce an electrical current.

The typical efficiencies of thermoelectric generators are around 8-10%. Older Seebeck-based devices used bimetallic junctions. More recent devices use semiconductor p-n junctions made from bismuth telluride (Bi2Te3), lead telluride (PbTe), calcium manganese oxide, Ge/SiGe superlattices [4].

These are solid state devices and unlike the previous ones have no moving parts. The choice of the material to be used for the fabrication depends also on the temperature. Figure 8 depicts a view of a TEG developed using bulk 2D Si/SiGe and Ge/SiGe superlattices, laterally patterned 1D nanowires and 0D quantum dots made from Ge/SiGe heterostructure technology.

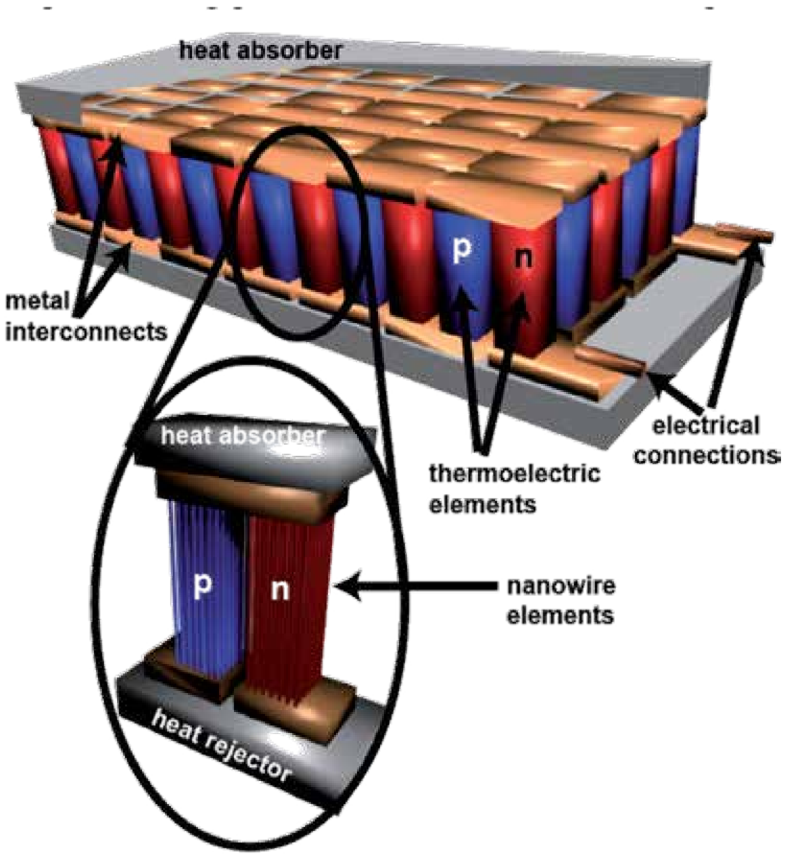


Figure 8. A Thermo-Electric Generator (from "Generate Renewable Energy Efficiently using Nanofabricated Silicon (GREEN Silicon)" project, EC FP7 ICT FET Proactive Initiative "Towards Zero Power ICT" Project No. 257750 University of Glasgow, U.K.; Politecnico di Milano, Italy; Universität Linz)

It is also possible to extract energy from radio waves. An antenna is used to convert an incident wave to a current. If the antenna is designed to work over a wide band, it is possible to extract a non negligible amount of energy. The transfer of power with radio waves is described by the Poynting vector *S*, a quantity representing the magnitude and direction of the flow of energy in electromagnetic waves.

The vector S is defined as the cross product $\vec{S} = (1/\mu)\vec{E} \times \vec{B}$, where μ is the permeability of the medium through which the radiation passes, *E* is the amplitude of the electric field, and *B* is the amplitude of the magnetic field. The direction of the vector $\vec{S}$ is perpendicular to the plane determined by the vectors $\vec{E}$ and $\vec{B}$. For a traveling electromagnetic wave, the Poynting vector points in the direction of the propagation of the wave. Given this, it is easy to understand that the power density of an electromagnetic wave is generally pretty low because E and B decrease with the square of the distance from the source. But if the harvester is close to the source or the source is very strong, for example close to a TV or radio broadcasting tower, it is possible to extract enough energy to power a sensor.

The current flowing from the antenna is an alternate current, so it needs to be rectified. The rectification process wastes energy and it must be taken into account when designing a harvesting system. As shown in Figure 9 a small signal diode can be placed in the center part of a dipole to rectify the RF signal.

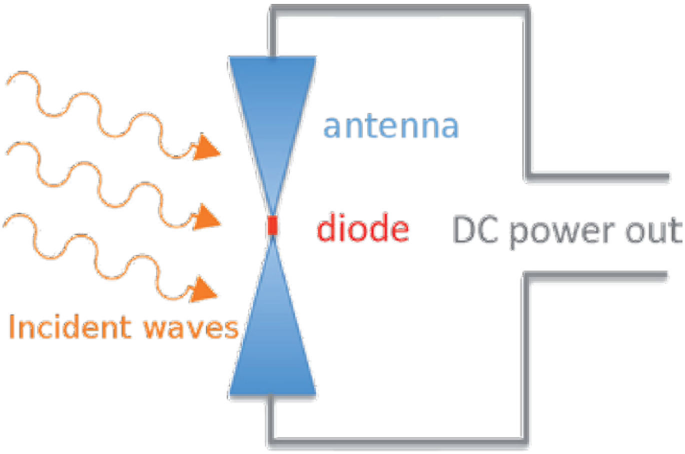


Figure 9. An antenna used for RF energy harvesting

Rectification can be made with diodes, like p-n or Schottky diodes, or active rectifiers. The first are passive devices and they introduce losses due to the threshold voltage of the junction. The latter are active devices and can reach higher efficiency in the energy conversion thanks to very low threshold voltages. They use active diodes, generally made with FET transistors, and require a control circuitry to turn them ON or OFF in the right sequence. So the choice of the rectifier is a non trivial task and must be taken into account when designing a harvesting system.

3.2. Power management

The energy coming from a harvester is not always constant in amplitude. Only few devices generate constant amplitude vibrations, for example a motor rotating at constant speed. The vibrations of a car, for example, are generally variable in amplitude and frequency. It can happen to have a very high spike and few millisecond later practically nothing.

In Figure 10 it is shown the time series of the real vibration of a wheel axle of a car: as it can be seen it is practically a random signal.

Assuming the use of a piezoelectric vibration energy harvester, the output voltage from the generator will be proportional to the amplitude of the accelerations. In some situations it will be possible to have peak voltage as high as 20 V or more. Consequently a voltage regulator will be needed to regulate the power supply voltage to a fixed value, for example at 3.3 V.

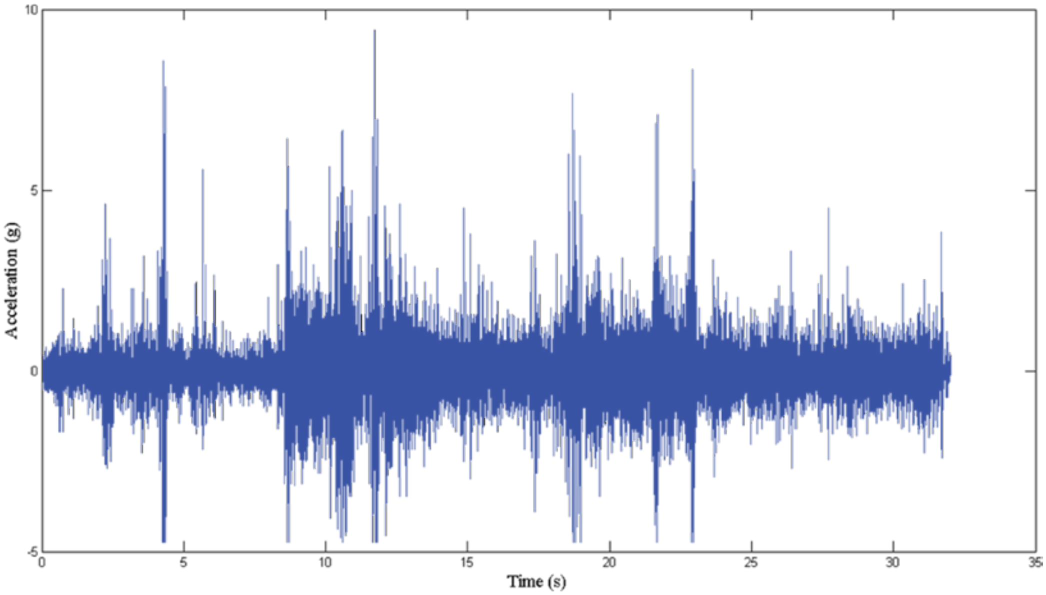


Figure 10. Time series of the vibration of a wheel axle

There are many types of voltage regulator, from a simple combination of a resistor and a Zener diode to the more complicated switching regulator. The first one has a low efficiency: a lot of energy is wasted in the resistor and in the junction when the voltage reaches the Zener threshold.

On the other side, switching regulators guarantee a higher efficiency in the regulation but they require working conditions a little bit more stringent: for example they require a short time, even if not null, to start working. Given that real vibrations may not be constant, it can happen that this kind of regulator will not start properly because the voltage coming from the harvester will sometimes be high enough to let it start but not always.

In some applications a simple and low-cost low dropout (LDO) voltage regulator can be the optimal solution. It is able to work with non constant voltage, it does not require any clock to work, and it guarantees a low dropout voltage, reducing the energy wasted due to voltage losses between its input and output.

A general scheme of the power supply chain can be represented as in Figure 8.The output of the voltage regulator is generally connected to an energy storage device. It can be a very low loss capacitor or a battery. The most important thing is to use a low self discharge and a very low internal impedance device. With a low self discharge the amount of energy wasted is small and a longer operating life can be guaranteed.

The low internal impedance is needed when the load requires high current peak during its working cycle. Small and high efficient thin film batteries are not suitable to work with high current peak because they are generally able to supply only 10 μA/cm^2 [5]. Using a 1000 μF tantalum capacitor as energy storage device, it is possible to evaluate the time required by the system in Figure 11 to reach the nominal voltage of 3.3 V. The vibrations used for the test are those represented in Figure 10. A double layer piezoelectric non-linear bi-stable energy harvester is used. The size of the cantilever used for the tests is 2.74 x 0.67 x 0.032 inches. The result, as shown in Figure 12, is a voltage across the capacitor rising from zero to 3.3 V in around 90 s with a non-constant slope, depending on the amplitude of the accelerations. In the same figure it is possible to see the instant in which the sensor is turned ON. After 56.4 s the voltage across the capacitor goes over the 2.35 V threshold and the voltage supervisor connects the capacitor to the load. In this way it is powered only when the supply voltage is high enough to guarantee the right supply voltage required by the electronic device.

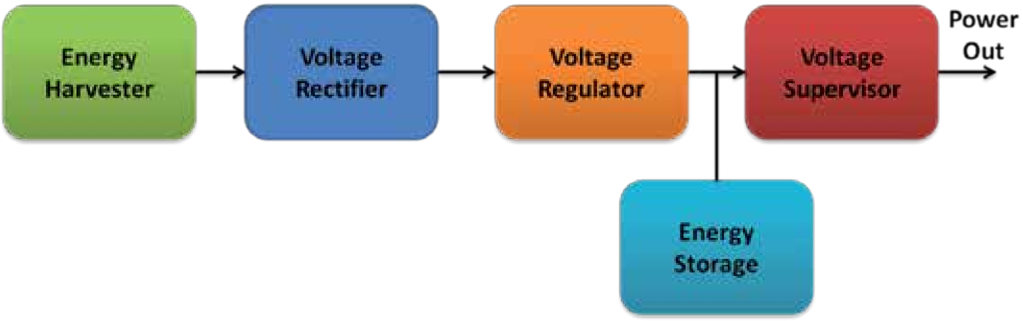


Figure 11. Power supply chain

3.3. Microcontroller

A sensor is a device that has to sense some parameters and then send their electrical representation for collection and different uses, depending on the application. A lot of different environmental variables can be acquired, processed and sent to a receiver. All these tasks are generally demanded by a microcontroller.

A microcontroller can be a very simple device, similar to a combinational logic network, or something more complex like a small computer with several peripherals. Microcontrollers are not microprocessors, they are like a small complex system composed by a CPU, a program memory, a data memory, timers, analog to digital converters, serial port interfaces, digital I/O etc. They are programmable: this is very important because they can do a lot of different tasks simply by changing the code and not the hardware.

All these components require to be powered. Generally an autonomous system is supposed to work for a very long time without maintenance. Each component has to be designed to work with the lowest power possible. In this way it is possible to obtain an extremely low power system suitable for powering with only the amount of energy extracted from the environment.

Some of today's microcontrollers are designed and realized bearing in mind very long lasting applications. Some companies have in their portfolio several devices with very low power consumption during the sleep operational mode. To reduce the overall power consumption, in fact, a microcontroller has to remain in sleep mode as long as it can.

Assuming that we acquire the temperature of a room once every ten seconds, it is possible to program our sensor to stay in sleep mode most of the time and in the active mode only for the time needed to sense the temperature, to prepare the data for the transmission and to send them through a radio frequency transceiver. In this way the power consumption over the 10 seconds period is the sum of the power required during the sleep and the active mode. Given that the active mode can last around 7 ms, it is clear that it is very important to have very low power consumption during the sleep time (9.993 s).

As an example, a small hybrid autonomous sensor has been developed and tested [6]. It uses a piezoelectric non-linear bi-stable energy harvester and two small solar cells to power a wireless node composed of a microcontroller of the MSP430 family and a RF transceiver produced by Texas Instruments. In sleep mode its current requirement is around 500 nA at 3.3 V. During the transmission of the data via radio, the current rises around to 25 mA at 3.3 V. The mean value of the current during the active mode is around 7 mA at 3.3 V.

It is possible to evaluate the energy required every 10 seconds as shown in Equation (1-3).

$$E_{sleep}=0.5*10^{-6}*3.3*9.993=16.488\ \mu J \tag{1}$$

$$E_{active}=7*10^{-3}*3.3*0.007=0.161\ mJ \tag{2}$$

$$E_{total}= E_{sleep}+E_{active}\cong E_{active} \tag{3}$$

It is clear that if the current required during the sleep time would be even only one order of magnitude higher, 5 μA, the energy required would be dominated by the sleep mode. In other words it would be possible to say that the greater part of the energy would be wasted into heat.

Programming the microcontroller is important too because the time required to stay in the active mode is proportional to the number of cycles the CPU has to perform. Generally microcontrollers are programmed in C, a general purpose programming language, or in other higher level languages; sometimes, to avoid the overhead of high level languages, it is better to write in assembly language to optimize the length of each function or routine.

3.4. Radio frequency transceiver

Once the data are in the memory of the microcontroller, they have to be transmitted to a receiver that will simply receive and store them in a database and eventually use them for some processing of automatic control or human activity. The transceiver by which the data are sent has the function of representing the data with radiofrequency signals and to transmit them sequentially over a certain distance.

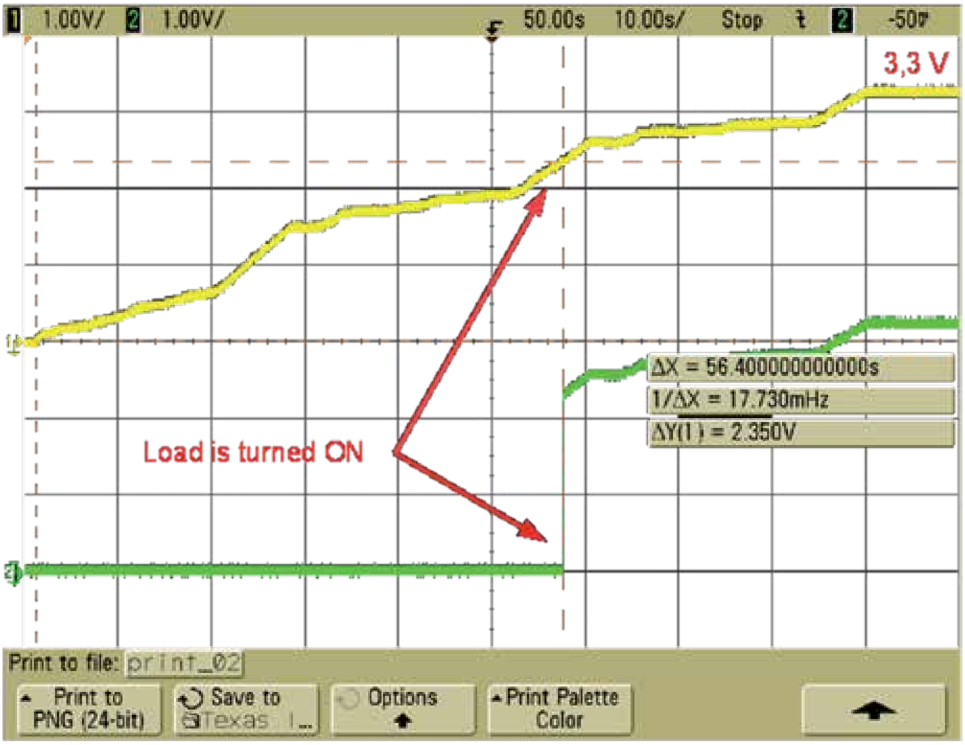


Figure 12. Voltage across a capacitor and output voltage from the supervisor

The wireless link can be in the range of few meters to several kilometers and even longer, but generally autonomous sensors are low power devices and the amount of power available for the radio transmission is of a few mW. Hence, they typically operate up to 100 meters.

The choice of the type of transceiver to be used is very significant and it depends on many variables. First of all transceivers distinguish one from each other from the operational frequency band. In each country there are some free frequencies dedicated to low power radio services, like telemetry, called ISM bands – Industrial, Scientific and Medical: no licenses are required to use these frequencies. The ISM frequency bands are summarised in Table 1.

One of the commonly used frequency bands is the 2.4 – 2.5 GHz. It is used for WiFi and Bluetooth communications. These are low power and high data rate communication technologies, generally not suitable for autonomous sensors because of their high computational cost; for example, their protocols require long time to establish a connection between two nodes.

Autonomous sensors are devices that generally transmit few bytes of data and the communications last few milliseconds. There are some other technologies that can be used for this purpose. Completely proprietary protocols can be implemented over a radio channel using modulation schemes like OOK, FSK or QPSK. These protocols can provide simple peer-to-peer communication or more complex network capabilities like routing and path optimization. But

Frequency range		Bandwidth	Center freq.	Availability
6.765 MHz	6.795 MHz	30 kHz	6.780 MHz	Subject to local acceptance
13.553 MHz	13.567 MHz	14 kHz	13.560 MHz	
26.957 MHz	27.283 MHz	326 kHz	27.120 MHz	
40.660 MHz	40.700 MHz	40 kHz	40.680 MHz	
433.050 MHz	434.790 MHz	1.74 MHz	433.920 MHz	Region 1 only and subject to local acceptance
902.000 MHz	928.000 MHz	26 MHz	915.000 MHz	Region 2 only
2.400 GHz	2.500 GHz	100 MHz	2.450 GHz	
5.725 GHz	5.875 GHz	150 MHz	5.800 GHz	
24.000 GHz	24.250 GHz	250 MHz	24.125 GHz	
61.000 GHz	61.500 GHz	500 MHz	61.250 GHz	Subject to local acceptance
122.000 GHz	123.000 GHz	1 GHz	122.500 GHz	Subject to local acceptance
244.000 GHz	246.000 GHz	2 GHz	245.000 GHz	Subject to local acceptance

Region 1 comprises: Europe, Africa, the Middle East west of the Persian Gulf including Iraq, the former Soviet Union and Mongolia.

Region 2 covers the Americas, Greenland and some of the eastern Pacific Islands.

Table 1. ISM (Industrial, Scientific and Medical) frequency band

everything has a computational cost, so generally it is preferable to work with simple peer-to-peer network with one access point, or coordinator, and several reduced functionalities nodes (star topology): refer to Figure 13.

As already discussed, the peak power required by a radio frequency transceiver can be very high, especially if compared with the power requested by the entire sensor. It must be taken into account that today's technology gives us the possibility to choose among many different modulation schemes. Each one has its own pros and cons: cost, complexity, bandwidth, spectral efficiency, signal to noise ratio (S/N) required at a given BER – Bit Error Rate (or Symbol Error Rate). The last one is a very important parameter that has to be taken into account when setting up a radio link because, knowing the amount of energy needed at the receiver to obtain a S/N ratio for a given probability of error and a given transmitting distance, it is possible to set the level of the transmitted power.

Figure 14 depicts the BER as a function of the ratio between the energy per bit E_0 and the noise N_0 for three different modulations techniques. If the required BER is 10^{-3}, for a PSK receiver the ratio E_0/N_0 must be a little bit less than 10 dB. If the distance between the transmitter and the receiver remains the same, to obtain the same probability of error on the bit (BER = 10^{-3}) using a FSK or ASK receiver a higher power will be required. Looking at the graph in Figure 14 it is possible to evaluate the required E_0/N_0 ratio for the desired modulation: it is a little bit less than 16 dB, 6 dB more than for the PSK modulation.

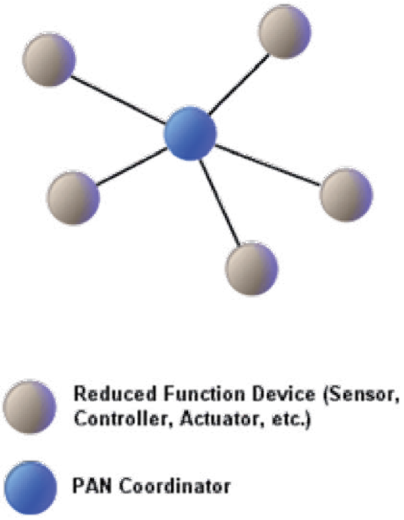

Reduced Function Device (Sensor, Controller, Actuator, etc.)
PAN Coordinator

Figure 13. Star topology network

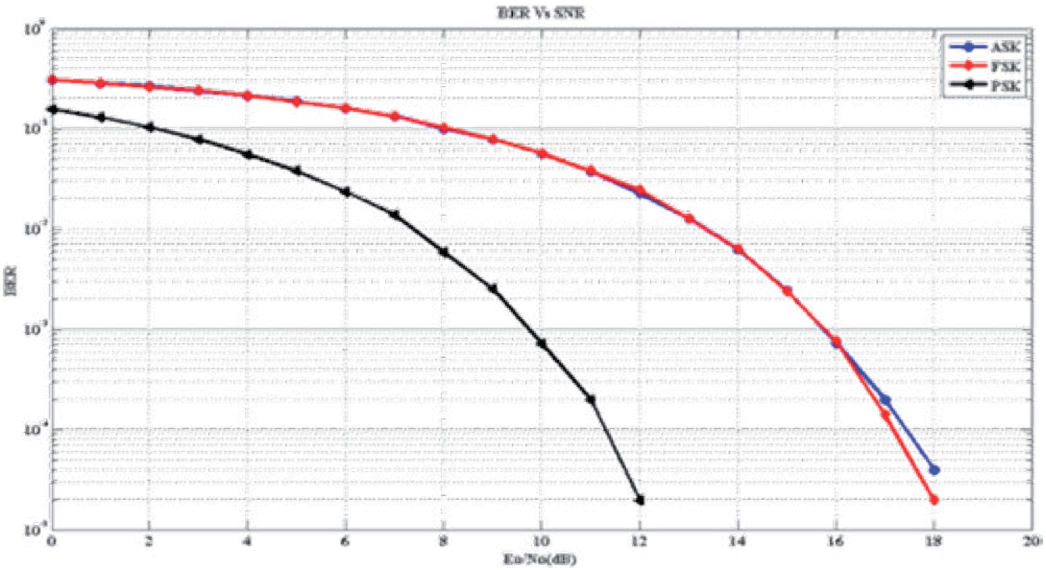


Figure 14. BER vs SNR for three different comment modulations

Last but not least the use of a high performance antenna in terms of gain and radiation pattern, is desired because it will reduce the amount of power required to cover the distance between the transmitter and the receiver. Several types of antennas have been designed and realized. There is not a perfect antenna for every application, so each time the right antenna has to be selected to obtain the desired performance.

Antennas can be printed on a circuit or can be simply realized as a standalone device. They can also be realized as a small integrated chip, for example in LTCC technology, or directly printed on a dielectric substrate with an ink-jet printer.

4. RFID based sensor

Autonomous sensors may sometimes be interrogated contactless without the use of battery. This type of devices is interesting for applications in which the use of wired solutions, for example in-package or in body measurements, would be very difficult. For these applications, the batteries maintenance and replacement is always a problem.

In literature there are reported applications [7,8] in which the autonomous sensing system uses the electromagnetic field to power the device. In general RFID tags are divided into two main categories: passive; they derive their operational energy from the RFID reader signal, do not have a real transmitter, modulate and irradiate with their antenna the signal transmitted by the reader. The distances at which they can operate are, at most, of the order of a few meters or a few centimeters depending on the operational frequency. Active; these are powered by batteries. They incorporate both receiver and transmitter as the reader. They have usually large memory, often rewritable. The distances at which they can operate depends on the transmitter and batteries and typically are, at most, of the order of 200 meters.

From the power consumption prospect, RFID-enabled sensors can also be divided into two categories: active and passive ones. The active RFID-enabled sensor tags use batteries to power their communication circuitry, and benefit from relatively long wireless range. However, the need of external battery limits their applications only to where battery replacements are possible and affordable. In the case of passive RFID-enabled sensor tags, when the reader interrogates the passive sensor system, the transmitted RF signal is used to power the system and then the sensing data are communicated to the reader wirelessly through the electromagnetic field and an antenna interface. In fact the increasing development of sensors based on RFID technology, is built on a combination of antennas coils and IC using the principle of backscattering to enable digital and analog sensing capability.

Recently, the interest for materials used in sensing applications is growing: an ultra sensitive composite which can be printed directly on the same substrate together with the antenna, when inkjet printing technology is chosen for an ultra fast prototyping production and for low cost, flexible solutions, are being analyzed. Carbon Nanotubes (CNT) composites have been found to have electrical conductance highly sensitive to extremely small quantities of gases. In this section we aim to give an example as a proof of concept, of a novel approach that contemplates the elimination of the IC chip and its replacement with a sensing element directly integrated with the antenna for the transmission of sensitive data wirelessly in a near "zero-power" fashion.

In [9] a CNT (Carbon Nanotubes)-based RFID-enabled sensor node was presented as a proof of concept. The sensitive element was chosen to be a film made by CNTs layers. The electrical resistance of CNTs film varies depending on the presence of gasses in the surrounding environment. As mentioned before, in passive RFIDs the reader sends an interrogation signal to the RFID tag which is formed by an antenna and an IC chip as load. The integrated circuit responds to the reader varying its input impedance, thus modulating the signal response back to the reader. Usually the type of modulation used is the ASK modulation, amplitude shift keying, in which the integrated circuit varies its impedance between matched and mismatched state.

In an RFID system, the antenna reflection coefficient can be calculated to evaluate the reflected wave strength. The same mechanism can be used to realize RFID-enabled sensor nodes. The SWNT film works as a tuneable resistor with a value that depends on the existence of the gas under test in the surrounding environment. The RFID reader monitors the backscattered power level and when there are changes in this level, it means that the target gas is detected. The conceptual diagram of the working principle is shown in Figure 15.

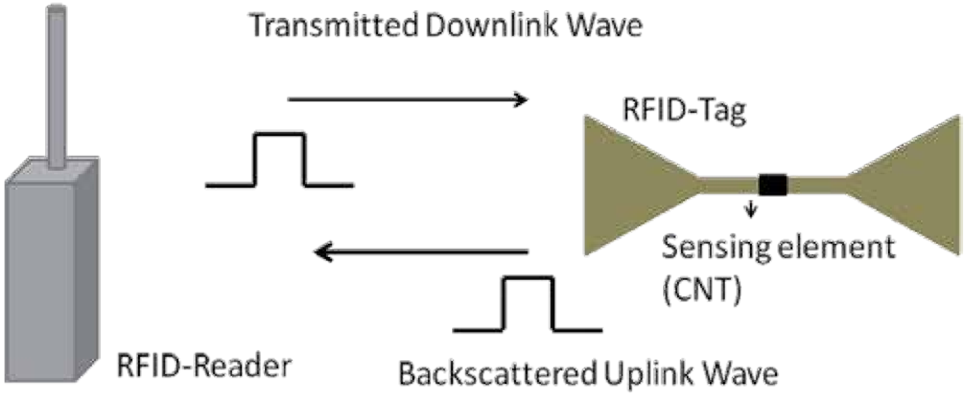


Figure 15. Conceptual diagram of the proposed RFID-enabled sensor module.

5. Conclusions

In this chapter an overview of autonomous sensors has been presented. These devices are powered with just the energy available from the environment, thus eliminating the need of replacing batteries. The required energy is made available by many different types of energy harvesters. Each one differs from the other by the kind of energy that is able to convert.

In section 2 a diagram of a basic sensor has been presented. Its fundamental parts have been highlighted and briefly described trying to focus the attention to the power requirement theme.

In section 3 the energy harvesting topic has been discussed focusing the attention on the basic principles behind each technology. There are some different approaches and each one has its own advantages and disadvantages.

Subsequently the topic of the power management has been presented and the chain of power conditioning analyzed. It has been shown that three fundamental parts are required: a voltage regulator, a voltage supervisor and an energy storage device.

Then the active circuitry of the sensor has been described. It is generally composed by a microcontroller, some sensing devices and a radio frequency transceiver. Many different devices are already on the market but just few are suitable for energy harvesting applications because of their power requirements. In fact it is very important to use extreme low power devices because the amount of energy available from a typical energy harvester of centimeters scale is orders of magnitude lower than the one available from a common AAA battery.

A brief analysis has followed about the modulation schemes regarding radio communication. When choosing a wireless transceiver, it is important to take into account the amount of energy required to transmit each bit of information to have a given probability of error at the receiver side. It would be desirable to have the lower power transmitted with the lower error rate: this generally means more complicated modulations and, consequently, a more complicated circuitry of the radio.

Finally, a short description of the RFID passive devices has been discussed. These sensors obtain their own required energy from an electromagnetic field radiated by a reader. In this way they can be considered truly autonomous systems because they do not use batteries and they comprise of all the circuitries required to convert the energy from the EM waves, to acquire data and to transmit them to the receiver.

Author details

Francesco Orfei* and Giulia Orecchini

*Address all correspondence to: francesco.orfei@nipslab.org

Nips Laboratory, Department of Physics, University of Perugia, Perugia, Italy

References

[1] Dargie, W. and Poellabauer, C., Fundamentals of wireless sensor networks: theory and practice, John Wiley and Sons, 2010 ISBN 978-0-470-99765-9.

[2] Luca Gammaitoni, There's plenty of energy at the bottom (Micro and nano scale nonlinear noise harvesting), Contemporary Physics Vol 53 (2) Pages 119-135, 2012.

[3] F. Cottone, H. Vocca, L. Gammaitoni, "Nonlinear Energy Harvesting" Phys. Rev. Lett. 102, 080601 (2009)

[4] S.C. Cecchi, T. Etzelstorfer, E. Müller, D. Chrastina, G. Isella, J. Stangl, A. Samarelli, L. Ferre Llin and D.J. Paul, "Ge/SiGe superlattices for thermoelectric energy conversion devices" Journal of Materials Science (Accepted for publication 2012) - doi 10.1007/s10853-012-6825-0

[5] Solid state thin-film lithium battery systems, N.J. Dudney, B.J. Neudecker, Solid State Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6030, USA

[6] Orfei, Francesco, Mincigrucci Riccardo, Neri Igor, Travasso Flavio, Vocca Helios, and Gammaitoni Luca, "Hybrid Autonomous Transceiver", EDERC2012: 5th European DSP Education and Research Conference, 13-14/09/2012, Amsterdam, (2012)

[7] E. L. Tan, W. N. Ng, R. Shao, B. D. Pereles, K. G. Ong, A wireless, passive sensor for quantifying packaged food quality, Sensors 2007,7 (2007),1747-1756.

[8] A. Vergara, E. Llobet, J.L. Ramrez, P. Ivanova, L. Fonseca, S. Zampolli, A. Scorzoni, T. Becker, S. Marco, J. Wollenstein, An RFID reader with onboard sensing capability for monitoring fruit quality, Sensors and Actuators B 127 (2007) 143–149.

[9] L. Yang, G. Orecchini, G. Shaker. H. Lee, M.M. Tentzeris, " Battery-free RFID enabled wireless sensor", Procs. of the 2010 IEEE-IMS Symposium, pp.1528 1531, Anaheim, California, May 2010.

Edited by Giorgos Fagas, Luca Gammaitoni,
Douglas Paul and Gabriel Abadal Berini

A sustainable future for our information society relies on bridging the gap between the energy required to operate portable ICT devices with the energy available from portable/mobile sources. The only viable solution is attacking the gap from both sides, i.e. to reduce the amount of energy dissipated during computation and to improve the efficiency in energy harvesting technologies. This requires deeper and broader knowledge of fundamental processes and thorough understanding of how they apply to materials and engineering at the nanoscale, all the way up to the design of energy-efficient electronics. This textbook is a first attempt to discuss such concepts towards Zero-Power ICT. The content is accessible to advanced undergraduates and early year researchers fascinated by this topic.

The book is realized through the EU-funded ZEROPOWER project.

Photo by NicoElNino / iStock

IntechOpen

ISBN 978-953-51-6358-9