STRUCTURED ELECTRONIC DESIGN

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

ANALOG CIRCUITS AND SIGNAL PROCESSING

Consulting Editor: **Mohammed Ismail**. *Ohio State University*

Related Titles:

HIGH SPEED A/D CONVERTERS: Understanding Data Converters Through SPICE
A. Moscovici
ISBN: 0-7923-7276-X

ANALOG TEST SIGNAL GENERATION USING PERIODIC -ENCODED DATA STREAMS
B. Dufort, G.W. Roberts
ISBN: 0-7923-7211-5

HIGH-ACCURACY CMOS SMART TEMPERATURE SENSORS
A. Bakker, J. Huijsing
ISBN: 0-7923-7217-4

DESIGN, SIMULATION AND APPLICATIONS OF INDUCTORS AND TRANSFORMERS FOR Si RF Ics
A.M. Niknejad, R.G. Meyer
ISBN: 0-7923-7986-1

DESIGN AND IMPLEMENTATION
B.E. Jonsson
ISBN: 0-7923-7871-7

RESEARCH PERSPECTIVES ON DYNAMIC TRANSLINEAR AND LOG-DOMAIN CIRCUITS
W.A. Serdijn, J. Mulder
ISBN: 0-7923-7811-3

CMOS DATA CONVERTERS FOR COMMUNICATIONS
M. Gustavsson, J. Wikner, N. Tan
ISBN: 0-7923-7780-X

DESIGN AND ANALYSIS OF INTEGRATOR-BASED LOG -DOMAIN FILTER CIRCUITS
G.W. Roberts, V. W. Leung
ISBN: 0-7923-8699-X

VISION CHIP
A. Moini
ISBN: 0-7923-8664-7

COMPACT LOW-VOLTAGE AND HIGH-SPEED CMOS, BiCMOS AND BIPOLAR OPERATIONAL AMPLIFIERS
K-J. de Langen, J. Huijsing
ISBN: 0-7923-8623-X

CONTINUOUS-TIME DELTA-SIGMA MODULATORS FOR HIGH-SPEED A/D CONVERTERS: Theory, Practice and Fundamental Performance Limits
J.A. Cherry, W. M. Snelgrove
ISBN: 0-7923-8625-6

LEARNING ON SILICON: Adaptive VLSI Neural Systems
G. Cauwenberghs, M.A. Bayoumi
ISBN: 0-7923-8555-1

ANALOG LAYOUT GENERATION FOR PERFORMANCE AND MANUFACTURABILITY
K. Larnpaert, G. Gielen, W. Sansen
ISBN: 0-7923-8479-2

CMOS CURRENT AMPLIFIERS
G. Palmisano, G. Palumbo, S. Pennisi
ISBN: 0-7923-8469-5

STRUCTURED ELECTRONIC DESIGN

High-Performance Harmonic Oscillators and Bandgap References

by

Arie van Staveren
Chris J.M. Verhoeven
Delft University of Technology, The Netherlands

and

Arthur H.M. van Roermund
Eindhoven University, The Netherlands

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-48169-3
Print ISBN: 0-7923-7283-2

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

Contents

Preface

Analog design has compared with digital design, still a flavor of art. Art can be beautiful. However, art in itself is difficult to teach to students as well as to transfer from experienced analog designers to new trainee designers in companies.

At the Electronics Research Laboratory of the Delft University of Technology for many years the structuring and systematizing of analog design has been studied. Such studies result in, besides new insights and new circuits, an ordering of knowledge that is already known, i.e. a classification of circuit solutions and methods is obtained. In this way a designer is enabled to get an overview of (a part of) the analog design field. He or she doesn't need to know by heart a long list of circuit solutions and methods. A systematic classification which clearly shows the trade-offs gives ordering and relative performance of solutions. This speeds up the design process very much. Further, by classification, solutions not known so far can be found (invented). This is because a new circuit solution is indicated by an empty class of solutions.

In this book the structured electronic design of high performance harmonic oscillators and bandgap references is described. The book can be used by experienced engineers and researchers but also this material can be well applied for advanced courses in analog design.

Arie van Staveren, October 2000

Acknowledgements

This book has appeared before as a Ph.D. thesis of the Delft University of Technology. I am grateful to everyone who has helped me during my research and writing. Mentioning all the names is impractical and I have therefore restricted myself to those people who contributed in one way or another to a substantial portion of this work.

Thanks go to the scientific staff of the Electronics Research Laboratory of the Delft University of Technology for creating the stimulating academic atmosphere. Specifically I like to mention the names of my promoters Dr.ir. E.H. Nordholt and prof.dr.ir. A.H.M. van Roermund, both at that time professor at our lab. Also special thanks go to Chris Verhoeven, my room-mate and co-promoter who taught me a lot with his very structured and dogmatic way of reasoning in the field of electronic design. He involved me at an early stage of my carrier in the very inspiring and challenging field of education.

My thanks also go to the Ph.D. students of our lab during my thesis work. Here, special thanks go to Bert Monna. For more than three years he was my room-mate and we had a lot of discussions on electronics and all of the things that make a part of our lives.

Many students have been a source of inspiration during my research. I would like to thank the students who did (a part) of their finals with me, these are, in alphabetical order: Mohamed Acem, Mohamed Aouichi, Joep Bosch, Frank Kuijstermans, Maikel Mardjan, Jeroen van Velzen and from Nicaragua: Pablo Vasquez Obando.

The technical staff also deserves thanks as they provide invaluable practical support. In particular, I would like to offer thanks to Rob Janse who drew the numerous pictures for this book, Wil Straver for the help in designing chips and testing them, Loek van Schie for taking care of the PCs and Jan Nusteling for maintaining the Unix computer system.

Writing in a language other than your native language makes the process of writing even more difficult. I would therefore like to thank Simon North for the corrections he made to the “English” I wrote.

Arie van Staveren, October 2000

Chapter 1

Introduction

For many years designers have been creating electronic circuits and systems for many types of applications. The designer has to translate requirements dictated by the application into an electronic circuit within the constraints of the available technology.

In the early days of the electronics the designer had to realize his circuit using only passive components and vacuum tubes. In those days the vacuum tube was a costly element and, consequently, the designer was faced with the problem of using as few of them as possible. The circuits were relatively small.

After the invention of the transistor by Bardeen at the end of 1947 [1], the designer was able to use more active components. He gained more freedom in his design process, but, as more became possible, more was asked from him. The circuits he had to design became more complex.

Nowadays, the designer is able to use numerous transistors in his circuits owing to the common availability of IC technology. However, his design problem has become even more complex. Ongoing miniaturization has driven the designer to realize larger systems on a chip. Physical sizes decrease, but, in contrast, the functionality of the systems increases.

The number of requirements which must be fulfilled by an electronic implementation of a system has been increased considerably. For instance, a low power consumption and a low supply voltage are more or less modern requirements. Besides more requirements, the required figures of merit a circuit must have, gradually increase.

The complete design problem is becoming too large to tackle without a structured design method. It is hard to keep track of the implications on all the design aspects when a single parameter is varied. A design process should be described in an objective language and orthogonalization should be one of the key words. The complete design process should be divided into several independent design steps. Each of these design steps should optimize a separate

design aspect of the total solution. When the design aspects used are completely orthogonal, the optimum solution is found. Otherwise, fine tuning could be required to end up at the optimum solution.

Outline of the Book

This book discusses and applies a structured design method for electronic circuits. The method presented orthogonalizes the design on basis of fundamental information-theoretical considerations. The cells by which the implementations are finally realized are single devices. Chapter 2 introduces the formal method. It describes the required signal transfer in the mathematical language as a differential equation, which makes the method independent of technology. Inspection of the differential equations leads to the distinction of homogeneous and inhomogeneous differential equations. Homogeneous differential equations do not have an excitation whereas inhomogeneous differential equations have. The corresponding electronic implementations are called homogeneous and inhomogeneous circuits, respectively. Consequently, circuits are divided into those with an input signal and those without. The design aspects to be orthogonalized in the design process are found from Shannon's theorem [2]: noise, bandwidth and signal power.

Linear homogeneous circuits have been chosen as the topic of this book. These are circuits that generate a reference signal without an input signal. For each order of differential equation the corresponding electronic functions are derived. From the first-order linear homogeneous differential equation the DC reference is obtained. The electronic equivalent of the second-order linear homogeneous differential equation is the harmonic frequency reference. Higher-order linear homogeneous differential equations do not result in other functions. An implementation of the DC reference is the bandgap reference. The harmonic oscillator is an implementation of the harmonic frequency reference. Further, it is argued in this chapter that the amplifier is an inevitable building block for realizing these homogeneous circuits and will be discussed, consequently.

Modern systems are increasingly realized in a portable fashion. One of the main contributors to the size and weight of portable systems are the batteries. Lowering the supply voltage leads to a reduction in the *number* of batteries, e.g. for a 1 V supply voltage only one battery is required. Lowering the power consumption results in the use of *smaller* batteries. Therefore, low-voltage low-power design is a hot topic nowadays, as then only one, relatively small, battery can be used. Chapter 3 discusses the impact of a low-voltage low-power constraint on the performance of circuits related to noise, bandwidth and signal power. In order to reduce the power consumption of the circuits, class-AB biasing is a very powerful method. Therefore, class-AB biasing is also dealt with in this chapter.

Amplifiers are required for the design of both, bandgap references and har-

monic oscillators. Therefore, the structured design method is first applied to the amplifier in Chapter 4. A lot is already known about how to design amplifiers [3], [4], [5], [6], [7] and [8]. In addition to the new theory which is presented, common knowledge from literature is used to give a more complete picture. Ideally, the amplifier is not limited in bandwidth and, consequently, does not alter the order of the differential equation describing the complete system. However, in practice, amplifiers do have speed limitations and may alter the order of the differential equation. In order to keep this influence small enough, the bandwidth of the amplifier must be large enough. In the literature [9], [10] and [11] the frequency behavior of amplifiers is mostly reduced to a first-order behavior. However, to obtain maximal bandwidth for a given amplifier, the largest number of poles possible has to be used. Chapter 4 presents a structured method for frequency compensation of the amplifier with the maximum use of the speed potentials of the devices used.

For oscillators, the phase noise is a very important design aspect. For high-performance oscillators this phase noise must be relatively low in order to have a very high frequency stability. At the cost of an increased power consumption, the relative distance of the phase-noise power to the carrier power can easily by increased by straightforwardly increasing the resonator power. A more convenient, i.e. power efficient, way to increase this carrier-to-noise ratio (CNR) is by optimizing the oscillator circuit such that the noise contribution of the, inevitable, active part is reduced to a minimum. In Chapter 5 it is shown that, among other things, tapping of a resonator is a very powerful method to reduce this contribution and consequently, increase the CNR.

Many authors have already dealt with the harmonic oscillator. However, each from another point of view or with another focal point. Boon [12] introduced orthogonality in the design of harmonic oscillators: the required undamping and the amplitude control are realized such that they can be designed independently of each other and, even more importantly, can function independently of each other. As orthogonality is also a key word in this book, the basic models for the harmonic oscillator as described by Boon [12] will be used as the basis for the discussions in Chapter 5.

Bandgap references have been designed for many years now [13] and [14]. However, a structured design method has never come close to being formulated. In this book the structured design method as presented in the following chapter is applied to the design of bandgap references in Chapter 6. Limits with respect to noise are derived and design examples are also given, including the best low-voltage, i.e. 1 V, bandgap reference published up to now.

Finally, Chapter 7 discusses and concludes the book.

Bibliography

[1] J. Bardeen and W.H. Brattain. The transistor, a semi-conductor triode. *Physical Review,* 74:230–231, June 1948.

[2] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal,* 27(3):379–432 and 623–656, July 1948.

[3] H.S. Black. Stabilized feedback amplifiers. *The Bell System Technical Journal,* pages 1–18, January 1934.

[4] H.W. Bode. *Network Analysis and Feedback Amplifier Design.* Van Nostrand, New York, 1945.

[5] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[6] Z.Y. Chang and W.M.C. Sansen. *Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies.* Kluwer Academic Publishers, Dordrecht, 1991.

[7] J. Davidse. *Analog Electronic Circuit Design.* Prentice Hall International (UK) Ltd, London, 1991.

[8] P.R. Gray and R.G. Meyer. *Analysis and Design of Analog Integrated Circuits.* John Wiley & Sons Inc., New York, 1993.

[9] J.E. Solomon. The monolithic op amp: A tutorial study. *IEEE Journal of Solid-State Circuits,* 9(6):314–332, December 1974.

[10] E.M. Cherry. A new result in negative-feedback theory, and its applications to audio amplifier design. *IEEE Journal of Circuit Theory and Applications,* 6(3):265–288, July 1978.

[11] P.R. Gray and R.G. Meyer. MOS operational amplifier design - a tutorial overview. *IEEE Journal of Solid-State Circuits,* 17(6):969–982, December 1982.

[12] C.A.M. Boon. *Design of High-Performance Negative-Feedback Oscillators.* PhD thesis, Delft University of Technology, September 1989.

[13] D.F. Hilbiber. A new semiconductor voltage standard. In *ISSCC Digest Technical Papers,* volume 7, pages 32–33, 1964.

[14] G.C.M. Meijer. *Integrated Circuits and Components for Bandgap References and Temperature Transducers.* PhD thesis, Delft University of Technology, March 1982.

Chapter 2

Structured electronic design

2.1 Introduction

The design of electronic circuits is becoming more difficult due to the larger number of requirements an implementation has to fulfill and due to the increasing complexity of the systems. To cope with this, the design process needs to be structured to be able to find a circuit in a reasonable time which meets all the specifications. The search for the best fitting circuit in the space of circuit solutions needs to be done in a structured way.

A structured design method can be achieved by, among other things, giving the design process a hierarchical structure. At the highest hierarchical level the required function is described. In the subsequent levels the implementation of this function is then done step by step, at each level obtaining a higher level of refinement of the models, et cetera.

Further, the optimization of the different design aspects is not done at once. Each design aspect is optimized independently of the other design aspects. Each separate optimization finds the solution space of order (n-1) in the total space of the n-th order for which the design is optimal with respect to that design aspect. The final optimum solution is found at the intersection point of the n solutions of order (n-1) representing the optimum for the n design aspects.

Structuring a design process implicitly means that the aim of the process is described in an efficient language. This language is required to accurately and unambiguously describe the specifications and to ease the evaluation of the circuit for whether it meets the specifications or not. The specifications for an electronic system, which are derived from an application, first of all describe the required *signal-processing function.* Besides this typical required signal-processing function, requirements are specified concerning the *quality aspects* of this signal-processing, for instance bandwidth and the allowed error due to the

spread in parameters. The language used to describe the design process should have the ability to make this distinction. The ideal signal-processing function has to be described at the start of the design process, and during the design process the quality aspects are introduced somehow.

2.2 Mathematical description language

The obvious language to describe the required signal-processing function is the mathematical description language. With this description language the required signal-processing function is described from the point of view of a high level of abstraction. It can be described independently of any form of implementation detail such as the topology, the power-supply voltage and the process to be used.

For mathematical problems, a large number of methods for solving them is available. For many decades mathematicians have been creating methods for optimizations, methods for manipulating expressions, et cetera. These methods are independent of any form of application: they are generally applicable. When a required signal-processing function is described in the mathematical description language, this complete tool box becomes a resource for the electronic designer in his design process. Especially, thanks to the high level of abstraction that is used in the mathematical language, and the independence of the mathematical methods of electronic design, new relations and methods can be found for electronic design and consequently, new solutions (circuits) might be found.

A lot of types of mathematical description languages are available. For instance, a filter transfer can be described by means of its pole-zero plot or by means of its state-space description. Further, a description language can include information about the sequence in which the operations are to be performed, the so-called algorithmic description languages. Computer languages are an example of this type. Languages only describing the required function are called functional description languages. These do not contain any information about the sequence of operations. This level of abstraction gives more freedom than the algorithmic languages. Therefore, a functional mathematical description language is chosen: differential equations.

As the functions to be implemented are now described by mathematical expressions, the design of electronic systems actually becomes a matter of mapping a mathematical function to silicon.

2.2.1 Mapping mathematics to silicon

A general mathematical way of describing systems is by means of differential equations. A differential equation relates the input of a system, the excitation, to the output, the response. The dependent variable of the differential equation

is the output of the system. For instance, the second-order linear differential equation given by:

$$a\frac{d^2e_o}{dt^2} + b\frac{de_o}{dt} + c = e_i \tag{2.1}$$

describes how the output quantity, e_o, depends on the input quantity, e_i.

2.2.1.1 Homogeneous and inhomogeneous differential equations

The complete solution of a differential equation is described by the combination of:

- a general solution;
- a particular solution.

The *general* solution of a differential equation is found for the case that the excitation is identical to zero, i.e. from the so-called homogeneous differential equation [1]. It describes the response of a system on an *initial state* in the case of a zero input signal.

The *particular* solution is found from the differential equation with an excitation, which corresponds to an inhomogeneous differential equation; it is found for a particular excitation. It describes the solution of the differential equation for this particular excitation. As the general solution for stable systems decays as a function of time, the particular solution describes the *steady state* output signal of a system as a result of an excitation. From this the input-to-output transfer can be derived for the system.

For electronic circuits, the corresponding distinction is made. They are divided into:

- homogeneous circuits;
- inhomogeneous circuits.

Homogeneous circuits are those circuits which do not have an input signal. Their output signal is a result of the initial state and the natural response of the circuit, analogous to the homogeneous differential equation and its general solution.

Inhomogeneous circuits are circuits which have an input signal. Their output signal is a result of this input signal, which can also be a steady state signal, and the steady state response of the circuit, analogous to the inhomogeneous differential equation and its particular solution for an input signal, describing the solution for infinite time.

[1] The term homogeneous is also used for differential equations fulfilling the constraint $f(\lambda x, \lambda y) = \lambda^n f(x, y)$. In that case the differential equation is said to be a homogeneous differential equation of the n-th order.

It should be noted that from the electronics point of view, circuits having no input signal are called autonomous circuits. However, from the mathematical point of view, autonomous differential equation are equations in which the independent variable does not appear explicitly, which is a completely different property of differential equations. As in this book the electronics are based on mathematics, the terms homogeneous and inhomogeneous are used.

2.2.1.2 Basic operators and operands

When functions to be implemented are described by differential equations, the required set of operators and operands is limited. Consider the non-linear differential equation as given by:

$$a * \frac{d^2y}{dt^2} + x * \frac{dy}{dt} - 4 * x/y = \int x dt + 6. \tag{2.2}$$

This differential equation comprises the following operators:

- differentiation $\frac{d..}{dt}$;
- integration $\int ..dt$;
- multiplication $..*..$;
- division ../..;
- addition .. + ..;
- subtraction .. − ..;
- equating .. = ...

The operands of these operators can be:

- a variable;
- a constant;
- a function.

The variable can be independent or dependent. An independent variable is the input signal, for instance. Its value varies over a complete range. When an independent variable is intended to have a constant value which may be different for several situations, it is often called a parameter. The independent parameter cannot be freely chosen. It is the output signal, for instance.

An operand may also be a function, sin(x) for instance, which results from another differential equation.

Besides these separate operators and operands, one combination of an operator and operand deserves special attention. This is:

- scaling $a * ...$

Scaling is a multiplication by a predefined constant. In contrast with the multiplication operator, scaling has only one input.

When for all these operators and operands a structured design method is available, other differential equations can be implemented straightforwardly.

In the case of very large differential equations, of a relatively high order, reduction of the order can be done as is commonly done in mathematics. Then, a set of lower-order differential equations is obtained which can be readily implemented as separate functional blocks. For several applications common functional blocks can be distinguished and as a consequence, the design of those blocks can be parameterized.

2.2.1.3 Basic functional blocks

From the distinction of homogeneous and inhomogeneous differential equations the analogous homogeneous and inhomogeneous circuits were defined. A similar kind of distinction can be made for the basic building blocks. They can be said to be homogeneous or inhomogeneous, or, when related to the signal-processing function that has to be realized with those blocks, the two different types of blocks can be identified:

- signal-processing blocks;
- signal-generating blocks.

Signal-processing blocks are filters, amplifiers, detectors and so on, whereas oscillators and bandgap references are examples of signal-generating blocks.

Restrictions are imposed with respect to the original differential equation as soon as the differential equation is replaced by a number of basic building blocks together with a precedence relation. This is because the precedence relation of the functional blocks introduces an algorithm which was not modeled in the differential equation. The introduction of the precedence relation changes the description from a functional description into an algorithmic description. For example, the block schematic of a radio receiver shown in figure 2.1 already assumes that the selectivity of the radio is realized completely after the mixer. However, structured design on the level of differential equations leads to the conclusion that selectivity is best made when the filtering is done before as well as after the mixer so that a higher dynamic range can be achieved [1].

In this book the structured design of homogeneous functional blocks is discussed. For those blocks the homogeneous solution or, in other words, its natural behavior is important. These blocks have no *signal* input port; the only input port they have is a power input port for the supply of power to a load.

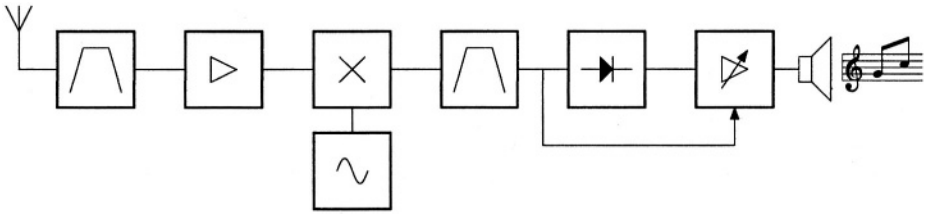

Figure 2.1: A block schematic of a radio receiver.

2.3 Electronic limitations

When using the mathematical description language for describing the function to be implemented (the required signal-processing function), the designer can keep all the functions exactly at this highest level. For instance, the multiplier in the radio, used to convert the desired frequency band to the intermediate frequency of the radio, multiplies an incoming frequency spectrum with the signal from a local oscillator:

$$e_{out} = e_{spectrum} * e_{localoscillator}. \tag{2.3}$$

The output signal is always the exact product of the two input signals. In practice, however, the speed of the mixer is limited. When relatively high-frequency signals are supplied to the input of the mixer, the output signal is, due to the limited speed, no longer the exact product of the two input signals. A non-ideal product operator is found.

At the highest level the designer is not bothered by these limitations of the practical implementations. In the subsequent levels these non-idealities can be introduced one by one. For the latter example, the time and space phenomena are not modeled in the mathematical description; one plus one is instantaneously two, no delay is found between excitation and solution. Whereas in the practical case one plus one has to be found, for instance, by adding two equal charges which are transported over some kind of channel with a non-zero length. As the speed of charge transport is limited, it will take some time before the solution is found, i.e. speed limitations.

At each level of the design trajectory new limitations are introduced. Step by step the models and circuits are refined to end up, finally, with a realistic circuit. In figure 2.2 a block diagram of the design trajectory from specification to physical realization is depicted. At the top, the application is found from which the specifications for the design are derived which must be fulfilled by the final realization. From these specifications a system description is generated, the description of the typical required signal-processing function and its quality. Subsequently, the system is divided into basic functional blocks, i.e. the lowering

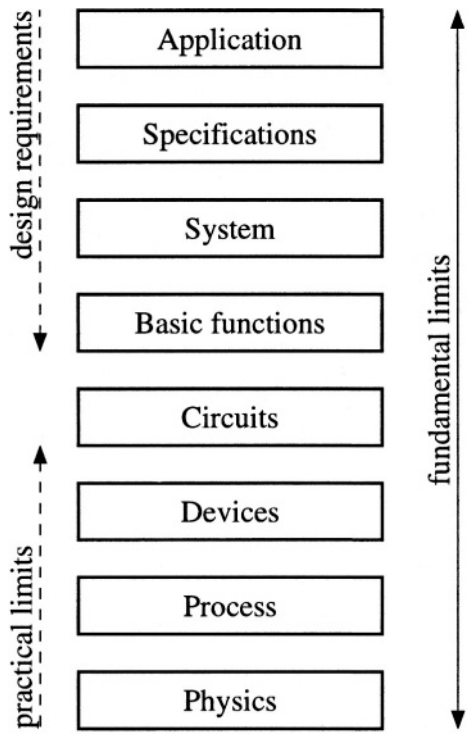

Figure 2.2: Block diagram of a design trajectory.

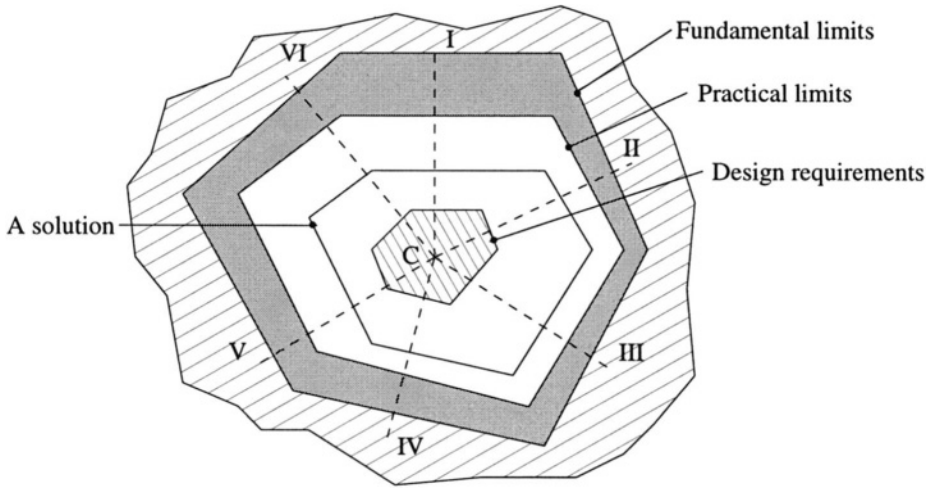

Figure 2.3: The space of circuit solutions.

of the order of the differential equations. These basic functional blocks are implemented by circuits. These circuits together are able to realize the required signal-processing function. The circuits, in turn, are constructed with devices which are realized in a specified process. This process is, finally, bounded by physical rules.

At each level, specific limits and requirements are found which have a certain hierarchy. The space of circuit solutions is depicted in 2.3 [2]. It should be noted that the multi-dimensional design process is depicted here in two dimensions only. In the direction of each axis (represented by dotted lines) the quality of a design aspect is assumed to be represented. Thus, the different quality aspects are assumed to be placed on a circular path. A solution is represented by a polygonal line. Of course, the number of sides of the solution line equals the number of design aspects in the design process. At the center (C) of this space, the circuits are found with the least performance. Going outwards, the performance of the circuits increases. When an axis from the center to the outer bounds is followed, three types of requirements and limits are found, see figure 2.4:

- design requirements;
- practical limits;
- fundamental limits.

These three types of boundary conditions for the design process are discussed in the following sections.

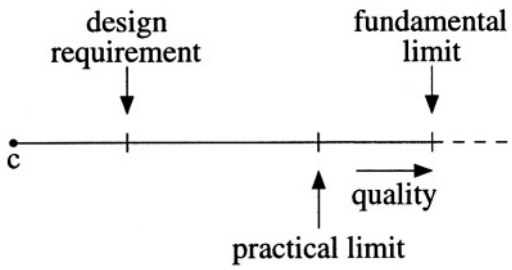

Figure 2.4: The quality of a design aspect related to requirements and limits.

2.3.1 Design requirements

The design requirements or specifications are determined by the application. They influence the design trajectory from the top side. This type of boundary condition sets the lower boundary of the system performance. The system must reach at *least* this quality.

Consequently, in figure 2.3, the conditions set by the design requirements are found at the inner part of the solution space. Going inside the area enclosed by these requirements results in a solution that is not good enough.

The design requirements derived from the application must reflect in an objective manner the principal required signal-processing function and its required quality. For instance, an application requires a frequency-reference signal with a frequency of 900 MHz which can be tuned in steps of 25 kHz. The specification: "An oscillator of 900 MHz which can be tuned in steps of 25 kHz is required", is an incorrect specification. The space of solutions is unnecessarily reduced by removing a frequency synthesizer from consideration.

The design requirements can have an influence on all the levels of the design trajectory. For instance, when a digital system is required, a totally different set of basic functions is found compared with an analog solution. Also the technology that is chosen in which the chip has to be realized is mostly different for digital circuits compared with the technology used for the integration of analog circuits. Thus at almost each level the design parameters are strongly influenced by the choice of the designer for digital instead of analog (or vice versa).

In contrast, introducing a low-voltage design requirement mainly alters the topology of the circuit. Instead of the voltage that is commonly used as the information carrying quantity, the current is often chosen. In this case mainly the implementation of the functional blocks might be influenced. When a charge pump is chosen to increase the supply voltage again, only this additional block is introduced and the design process remains almost unaltered.

2.3.2 Practical limits

The practical limitation are found to come mainly from the bottom side of the design trajectory. For a given technology the performance of a system cannot go beyond certain practical limits. For instance, the speed of the transistors in a certain technology is limited by the f_T of the process.

In a specific technology, the design aspects can reach *at most* the quality determined by these practical limits. Thus, the boundary set by these limitations is found at the outer part of the solution space.

The practical limitations can also affect all the levels of the design trajectory. For instance, the maximum power consumption per chip area of a process can influence the choice of the system concept. At the system level the number of operations can be minimized by means of mathematics. At circuit level this means that the transistors have to do less switching and thus less power is consumed. This type of power saving can be considerable [3]. Of course, at the other levels one also has to be careful with the power consumption to reach that low-power consumption.

2.3.3 Fundamental limits

In contrast to the design requirements and practical limits, which can vary from application to application, fundamental limits never change. The fundamental limits determine the ultimate performance that a design aspect can achieve. It is impossible to go beyond these limits. Therefore, these limits are found at the outermost border in the space of solutions. Clearly, the practical limits are always found inside the area enclosed by the fundamental limits.

The fundamental limitations mainly influence via the top and the bottom side of the design trajectory. From the bottom side, the speed of light in a vacuum is a fundamental limit. This is the highest speed that can be achieved. This maximum speed directly sets a limit on the speed by which information can be transported. As the information is modulated on a materialized carrier, molecules, electrons, et cetera, and the carrier cannot be faster than the speed of light, the information transport that is modulated on this carrier is also limited in its speed by the speed of light.

From the top side of the design trajectory, the fundamental limit defined by Shannon is found, for instance. He stated that the maximum information transporting capability of a channel, C, is given by [4]:

$$C = B \cdot^2 \log\left(1 + \frac{S}{N}\right), \tag{2.4}$$

where B is the bandwidth of the channel, S the maximum signal power that can be transported through the channel and N is the power level of the noise

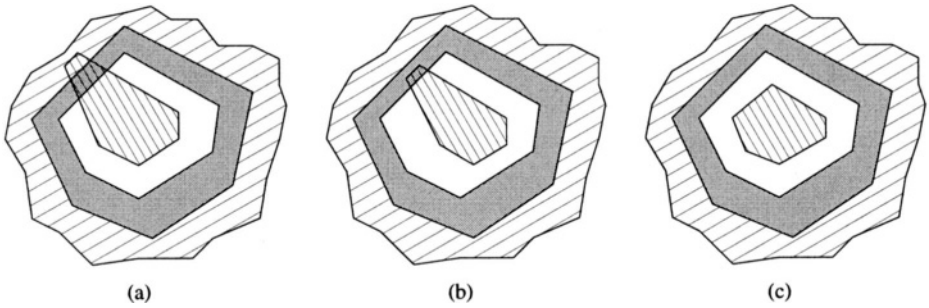

Figure 2.5: Three situations that are possible in the boundary conditions.

in the channel. If an attempt is made to transport more information through the channel than defined by this equation, information is definitely lost.

For the design process fundamental limits are important to know as they define what is maximally attainable. This is illustrated in figure 2.5. In figure 2.5a, the situation is depicted in which the design requirements are beyond the fundamental limits. In this case it is unnecessary to go any further in the design process as the design requirements are not feasible. The sooner this is detected, the more time and money is saved. Figure 2.5b, depicts the situation in which the design requirements are beyond the practical limits but within the fundamental limits. In this case the design is not fundamentally impossible. However, as the design requirements are beyond a practical limit, another technology, for instance, has to be used, for which this practical limit is beyond the design requirement. The third situation, depicted in figure 2.5c, illustrates the final situation when all the design requirements are within the practical limits. Now a practical solution exists.

Besides using the fundamental limits as a check for whether a solution exists or not, they can also be used to check the significance of a proposed system improvement. When a system is already close to the fundamental limit, it would probably be a waste of money and time to try to achieve that relatively small improvement.

When all the fundamental limits related to structured electronic design are classified, it appears that they are all related somehow to the fundamental limit as stated by Shannon. Consequently, signal power, noise power and bandwidth are the only fundamental limits concerned with in structured electronic design. It should be noted that chip area, power consumption, et cetera, are bounded by practical limits, i.e. the resources.

2.4 Optimization

The design of the circuit has now become the search for the optimum solution in the space of circuits which is limited by the three previously mentioned limitations. The optimum solution is a circuit that exactly meets all its design requirements, including the uncertainty introduced by process variation and so on, thus under worst-case conditions.

When the solution space is relatively small, the optimum solution can easily be found. However, the space is generally too large to be conquered without a structured search method, for instance, doing an exhaustive search which means that the complete solution space is searched solution by solution. A search method helps the designer to find in a structured and relatively fast way its nearby optimum solution or family of nearby optimal solutions.

2.4.1 Search methods

The starting point of the search path is often some sort of previous design. This design can be concrete, the previous version of the new circuit to be found or it can be more or less fictitious. In that case it can be some idea, for instance, from literature, or based on the experience of the designer.

The end point of the path is the optimum circuit or a family of optimal circuits. The search method indicates how the designer came from the starting point to this end point.

When classifying the search methods with respect to how the designer comes from the start to the end, three types are found. These search methods are:

- search by evolution;
- search by heuristics;
- search by creation.

These three methods are depicted in figure 2.6 and are discussed in the following sections.

2.4.1.1 Search by evolution

The new circuit is found by making relatively small changes in the previous circuit. In figure 2.6 this is indicated by method A. The variations made in the circuit are more or less based on rational considerations. The circuit is varied and when a small change in a specific direction results in an improvement of a design aspect, the circuit is tuned a bit more in that direction until the derivative of the improvement with respect to the parameter varied is zero.

The old circuit is tuned slightly to come to the improved solution. The new solution remains close to the previous one in the solution space. As it

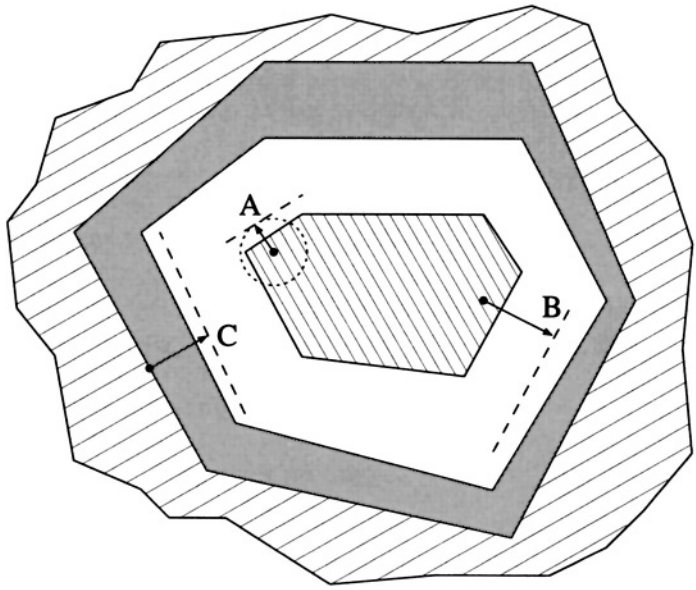

Figure 2.6: The three types of search methods. A) By evolution, B) by heuristics and C) by creation.

in principle uses only small steps, a large improvement of a circuit can cost a relatively large number of design iterations to come to the optimum solution or might not even be found because the iteration stuck at a local optimum.

2.4.1.2 Search by heuristics

This search method also uses the previous design as the starting point for the search method. But now the path to the new circuit is made by the intuition of the designer. On the basis of experience and ideas, he expects a relatively large improvement of the circuit when he changes the design in the solution space into a specific direction, see figure 2.6 method B. This method can give relatively large improvements in relatively short times. However, to come to the optimum solution, the designer must go further by evolution or possibly again by heuristics.

2.4.1.3 Search by creation

In contrast to the two previous methods, this search method does not use the previous design. The design is started from the ideal solution. As this ideal solution is on the fundamental boundary, it is very often an impractical solution.

When subsequently, step by step, the practical implementation is realized the nearest neighbor is found, see figure 2.6 method C.

The starting point of the search path is defined at the ideal solution. Subsequently, this ideal solution is refined by mathematical considerations to a realizable nearest neighbor. Ideally this search method does not use any other input from the designer than the design requirements. Therefore, this method is very suitable to be used in conjunction with the mathematical description language as this language also starts at the ideal solution and uses only objective criteria.

However, it may now seem that this method does not use the experience or knowledge of the designer, but this is not true. A long period of search by evolution and heuristics results in a lot of knowledge about and insight into designing high-performance circuits. Ordering and structuring all this knowledge and experience may result in a search method by creation, i.e. generalized knowledge. When some parts of the solution space were not already found by the designer, the structuring and ordering of knowledge helps in finding these areas and better solutions may be found. Thus designing by creation results from a relatively long history of designing by evolution and heuristics.

It should be noted that the method itself, used by search by creation, is changing due to evolution. New designs always differ in some respect from previous designs and may add, as a result, new knowledge to the design method.

2.4.2 Orthogonality

Design by creation starts with the ideal solution. This solution is the best with respect to all the fundamental criteria at the same time. This, implicitly means that one has to strive for orthogonality of the fundamental limits in the design method. Otherwise, weighting factors have to be introduced to be able to make the best compromise between two or more criteria. As weighting factors introduce subjectivity into the design process, they should be kept to a minimum. Orthogonality may help to put the remaining subjective weighting factors at the end of the design trajectory where the effect of the weighting can be evaluated for the complete design.

Orthogonality in the design process means that each design criterion can be optimized independently. In the solution space this means that each side of the polygonal line, representing a solution, can be moved independently of the other sides, see figure 2.7. By assuming that the design aspects are orthogonal, relatively simple expressions can be found that relate the practical limits, the design requirements and the fundamental limits for a design aspect. For instance, Groenewold [5] found that the maximum dynamic range, DR, of a filter is given by:

$$DR_{max} = \frac{V_{max}^2 C}{4kT\xi} \cdot f(H(j\omega)), \tag{2.5}$$

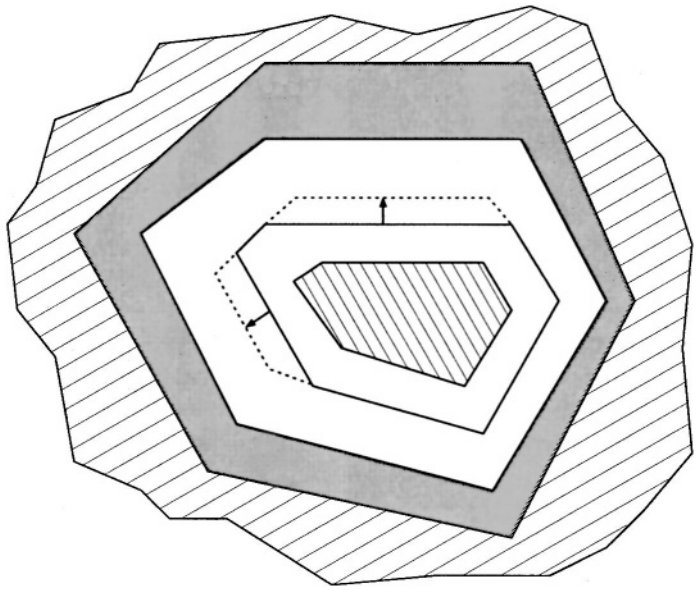

Figure 2.7: Orthogonality of the design aspects in the solution space.

where V_{max} is the highest possible voltage in the filter, C is the total capacitance used, ξ is the noise factor of the active components and $f(H(j\omega))$ is a function that depends only on the transfer function of the filter. This expression relates the fundamental aspects of channel capacity (DR) to practical aspects such as the total capacitance used capacitance (chip area), and noise factor of the amplifiers, and to design requirements such as the transfer function.

These kinds of expressions enable the designer to check, at a very early stage in the design process, where the design requirements are located in the solution space (cf. figure 2.5). Consequently, the designer can tell whether a solution is fundamentally and practically possible or not and whether a proposed technology is adequate or not.

2.5 Design strategy

The design process, as pointed out up to now, is rather abstract. In this section, the ideas presented previously are used to formulate a design strategy.

It makes sense to strive for orthogonality in the design process. The question that has to be answered is, what are the fundamental limitations in the circuit design that are considered to be orthogonal. The answer can be found when it is remembered what the primary function is that of electronic circuits in a

signal-processing system.

Electronic circuits in a signal-processing system are used, or are closely related, to transport and/or process information, which is picked up by whatever type of sensor, from the place where the source is, to the destination where the consumer is. This consumer may be interpreted in a very wide sense. It can be an antenna, a magnetic tape or a loudspeaker, for instance. Thus, the primary functions are the transportation and processing of information. This was already stated in the previous section, only in other words: the specifications for an electronic system, which are derived from an application, primarily describe the required *signal-processing function.*

The quality of the information transport has all to do with the capacity of the "channel". When the channel capacity is close to the quantity of information which has to be transported, the chance of an error occurring is greater than when the capacity of the channel is much larger than this quantity of information.

This channel capacity was shown, by Shannon [4], to be a fundamental quantity of a channel with respect to information transport. The channel capacity is related to noise power, signal power and bandwidth by equation (2.4), which is repeated here:

$$C = B \cdot^2 \log\left(1 + \frac{S}{N}\right).$$

The signal and noise are assumed to be *Gaussian distributed.* For noise sources with other characteristics, N must be replaced by the equivalent power which will always be lower than the actual power. Thus, with a Gaussian noise source a worst case is found. As, however, in electronic design, stochastic variables are mostly Gaussian distributed (thermal and shot noise, for instance), this measure is very readily applicable.

2.5.1 Noise

Noise, as it appears in Shannon's equation, describes the phenomena which add uncertain errors to the information transported through the channel. The information carrying signal can be altered by the channel in many different ways. For instance, additive noise just adds to the signal in contrast with parametric noise which influences the signal in a non-linear way through changing system parameters, like gain factors, et cetera [2].

The noise does not need to be small. Everything that reduces the entropy of the channel is called noise. When, for instance, the MSB of a digital system is

[2] The expression derived by Shannon for the channel capacity as given in equation (2.4), also assumes *additive* noise. When the noise is no longer additive but multiplicative, for instance, the principal integral function as given in [4] has to be evaluated for the channel capacity.

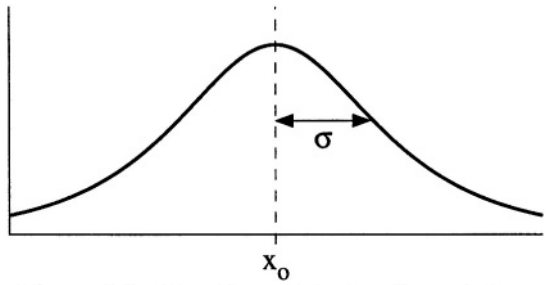

Figure 2.8: Modeling of stochastic variables.

not correctly functioning, the channel capacity is reduced as all the codes which use this MSB can no longer be used [3] However, the noise ís generally small.

A stochastic variable is often modeled as an expectation with a stochastic variation around it, see figure 2.8. The variation of the noisy signal around the expectation is modeled with the parameter σ; 68% of the time the signal has a value between $x_0 - \sigma$ and $x_0 + \sigma$. The standard deviation, σ, models the stochastic part and the expectation, x_0, the systematic part of the stochastic variable.

When the relation to the signal is taken into account, four different types of errors can be identified:

- signal-independent stochastic errors;
- signal-dependent stochastic errors;
- signal-independent systematic errors;
- signal-dependent systematic errors.

2.5.1.1 Signal-independent stochastic errors

This type of errors is, for instance, the thermal noise of resistors. This noise has the largest influence on the small signals in the channel. The noise level is related to the impedance level of the channel. To decrease the noise level with respect to the signal level, the impedance level of the system has to be lowered and therefore, more power has to be used.

2.5.1.2 Signal-dependent stochastic errors

This type of stochastic error is correlated to the signal transported through the channel. A very illustrative example is the noise behavior of a class B

[3] For this case equation (2.4) cannot be used straightforwardly as the noise is no longer independent of the signal. However, it is still additive. For the noise term, the noise entropy has to be used. This entropy is increased by the entropy of the signals using the MSB.

current-mode amplifier. For this moment only the shot noise of the active devices is considered. The power-density spectrum of the shot noise of active devices is proportional to the current flowing across their junctions. For the class B amplifier this current is directly related to the signal level and, consequently, the shot noise of the active devices depends on the signal level.

2.5.1.3 Signal-independent systematic errors

In principle, this type of errors does not need to reduce the capacity of a channel. This is easily seen when the basic definition of the capacity of a channel is remembered [4]:

> The channel capacity is the maximum entropy of the output signal minus the entropy of the errors added by the channel.

As the entropy of a deterministic signal is by definition zero, the channel capacity is not reduced. Of course, the maximum entropy must not be reduced by these systematic errors as otherwise a reduction of the capacity will still be observed. When the maximum entropy is reduced by a systematic error, the systematic error can be compensated such that the original situation is restored. It should be noted that this compensation introduces stochastic errors, but, these are generally relatively small.

An example of this type of error is the input offset voltage of an amplifier. The mean offset voltage is known and thus does not reduce the channel capacity due to additional uncertainties. However, when this offset voltage equals a base-emitter voltage and the power-supply voltage is only 1 V, the maximum possible signal level is reduced. This results in a reduction of the maximum channel entropy as stated before. When this input offset voltage is compensated such that the original maximum signal level is restored, the channel capacity is only slightly reduced by the stochastic errors of the compensating voltage.

2.5.1.4 Signal-dependent systematic errors

These errors are caused by the channel and depend on the momentary value of the signal. In contrast to the signal-dependent stochastic error, the relation between the momentary value of the errors and the signal is deterministic. In electronic design this type of error is often called non-linear distortion. Depending on the nature of the distortion, information may be lost. Therefore a distinction is made between:

- weak distortion;
- clipping distortion.

Weak distortion occurs when the static transfer of a system deviates from the intended transfer. The actual static transfer still has an inverse function so that a compensating function can be found for preventing the system from losing information. Thus, for instance, intentionally using the exponential transfer of a transistor does not mean that information is lost; by means of the inverse function, the logarithm, the information can be retrieved again.

The other type of distortion is found when the transfer of a system no longer has an inverse function, i.e. the transfer has become ambiguous. The original information can no longer be retrieved. This occurs, for instance, when signals clip to the supply voltage. Two different types of input signals, one causing clipping and the other being close to clipping, may result in the same output signal. A very straightforward example is the use of a limiter to get rid of all the amplitude information in a signal.

2.5.2 Signal power

The maximum signal power of a channel is a trivial limit. When the power of a signal exceeds the maximum possible power, information is lost by means of clipping, the signal-dependent systematic error which was discussed in the previous section. Of course, when the maximum possible signal level is increased, the capacity of the channel increases.

2.5.3 Dynamic range

Until now the noise (errors) and signal power have been examined independently of each other. The maximum signal power was shown to be limited by the resulting signal-dependent errors, especially clipping. This already indicates some relation between errors and signal. In Shannon's expression, the *ratio* of the noise power and the signal power is important. The channel capacity depends on the ratio of the maximum and minimum signal that can be transported. This ratio is called the dynamic range. In figure 2.9 the effect of a limited dynamic range on the entropy of the noise in the channel, $H(N)$, is depicted. For the moment consider the asymptote of the noise instead of the actual (dotted) function. For signal levels below S_{max} the entropy of the noise is given mainly by the signal-independent noise. Increasing the signal power, remaining below S_{max}, results in an increase of the channel capacity to C_{max}, at S_{max} by definition, because the entropy of S increases whereas the entropy of the noise remains constant. When increasing S any further and going beyond S_{max} the noise entropy increases causing the dynamic range to be limited to C_{max}. This noise entropy increases due to the increase of the signal-dependent errors, especially the systematic signal-dependent errors.

From figure 2.9 it follows that using signals beyond S_{max} results at least in a waste of power. The question now is, how S_{max} is defined. As can be seen,

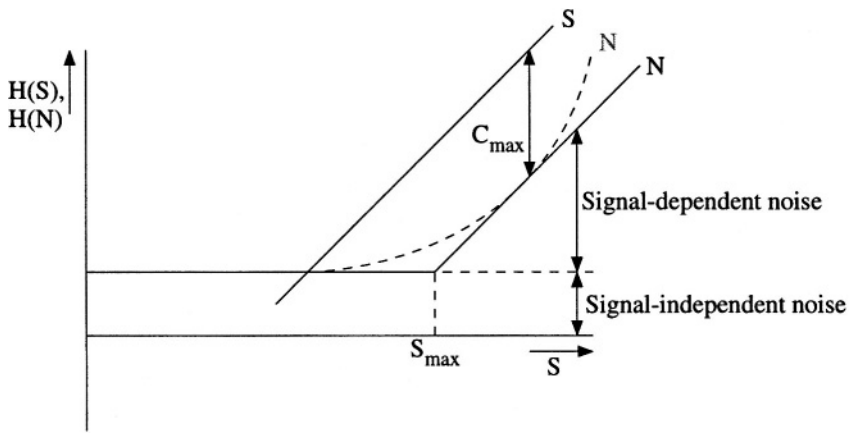

Figure 2.9: The effect of a limited dynamic range on the entropy of the noise, $H(N)$, in the channel.

S_{max} is at the intersection point of the two noise asymptotes. At this point the entropy of both noise types is equal. When they have the same statistical distributions their power is equal then. In that case S_{max} is found where the noise power is doubled due to the signal-dependent systematic noise.

Assuming equal statistical distribution is somewhat pessimistic as the noise floor is mostly Gaussian-distributed whereas the signal-dependent systematic noise is more likely to be non-Gaussian. In that case the ratio of the entropies is not equal to the ratio of the powers. As the entropy is maximal when a signal is Gaussian-distributed, the power of the signal-dependent systematic noise at S_{max}, the point of equal entropies for both noise types, is larger than the power corresponding to the noise floor. The exact ratio depends, of course, on the statistical behavior of the information and the transfer of the channel in the case of clipping.

It should be noted that the asymptotes in the figure only give an approximation. A realistic noise function can be as indicated by the dotted line in figure 2.9, resulting in an even higher S_{max} for maximal channel capacity. Also, for relatively high signal powers, the channel capacity is likely to decrease again.

2.5.4 Bandwidth

Signal power and noise power concern the maximum number of distinguishable signal levels. The maximum signal power sets the upper limit on the range of signal levels and the noise power sets the resolution. This description is independent of the way in which the information is coded such as, for instance, amplitude and frequency modulation.

Bandwidth concerns the speed of the channel. When a signal is too fast, it is not transported through the channel and information is lost.

When the bandwidth of a channel is doubled, the capacity is doubled. This bandwidth doubling, doubles the power consumption. When, in contrast, the channel capacity is doubled by changing the signal-to-noise ratio, the signal-to-noise ratio needs to be squared, assuming $S/N \gg 1$. For instance, for a signal-to-noise ratio of 100 this has to be increased to 10000. Thus, increasing the channel capacity is likely done the best, i.e. minimal additional required power, by increasing the bandwidth. As a consequence, when the bandwidth of a system is enlarged by exchanging it for signal-to-noise ratio such that the power consumption does not change, a net increase of channel capacity is likely to occur. For instance, when information is coded on a carrier by means of FM-modulation, the carrier is relatively insensitive to channel noise. This is a result of the larger bandwidth which is used, compared with the bandwidth of the information, so that a lower signal-to-noise ratio can be allowed in the channel to end up with the required capacity.

2.6 Fundamental limits for circuit design

Noise, signal power and bandwidth are the fundamental aspects of the channel capacity. Consequently, these need to be orthogonalized in circuit design in order to aim for the highest channel capacity, i.e. the ideal channel. How these aspects are related to circuit design is discussed next.

2.6.1 Noise

In a previous section, four types of errors were found. Two types having a probabilistic nature and two having a deterministic nature. The stochastic errors, fundamentally, limit the signal resolution. When signal details become smaller than this noise level, information is lost. The noise sources are mainly the thermal noise of resistors and the shot noise of active devices. Consequently, the noise in a circuit at the places where the signals are the smallest has to be minimized.

2.6.2 Signal power

For relatively large signals, practical circuits deviate from their intended static transfer and cause distortion. This distortion should be below a certain specified level. Information is lost due to distortion when signal information falls outside the frequency band of interest. In the case of clipping to a supply rail, the signal is terribly distorted and information is also lost. In both cases the distortion

can be said to be signal-*dependent* systematic errors. Therefore, this error type is a design issue for the maximization of the signal power.

The signal-*independent* systematic errors can be seen, at on hand, as a design issue for the noise minimization. Compensating for systematic errors introduces additional stochastic errors. However, when trimming is used, for instance in the case of an offset in a gain factor, the stochastic errors do not need to increase.

On the other hand, an offset voltage of 0.7 V of a base-emitter junction, limits the maximum signal when the supply voltage is only 1 V. Then the signal-independent systematic errors are a design issue for signal maximization. As low-voltage design is a hot topic nowadays, this type of error becomes more and more an issue for signal maximization.

2.6.3 Dynamic range

The dynamic range specifies the signal power and noise power as a relative figure, just as it appears in the expression given by Shannon. It takes the exchange of noise power and signal power into account. When the absolute signal levels can be freely chosen in a system, the dynamic range becomes the design issue. The designer can then use the signal levels for which the dynamic range is reached the best, i.e. the lowest power consumption is found. This is easily seen from the following example.

Consider the dynamic range of the collector current of a bipolar transistor. The shot noise is proportional to the collector current. The signal power, however, is proportional to the square of the collector current. As a consequence, the dynamic range of the bipolar transistor is proportional to the power consumption. The ongoing lowering of the power consumption of circuits nowadays leads to an inherent reduction in the dynamic range of those circuits.

To increase a circuit's dynamic range virtually, automatic gain control can be used. The gain of a system block is made dependent on the signal level. The amplification for small signals is larger than for large signals. The amplified small signals are less influenced by the noise of the succeeding system block. After processing by that system block the inverse action is done, i.e. the attenuation factor for relatively small signals is larger than for relatively large signals. The influence on the channel capacity of automatic gain control is depicted in figure 2.10. In figure 2.10a the dynamic range is depicted for a system without an automatic gain control. The noise level is independent of the signal level and thus the dynamic range is as indicated by the arrow. Figure 2.10b depicts the dynamic range in the case of automatic gain control. The noise level now depends on the signal level. For smaller signals the influence of the noise reduces. Different strategies can be used for controlling the gain in order to reduce this influence. Here it is assumed that the gain depends on the signal level over the complete range of signal levels. The real maximal DR is found for the highest signal level and can still be equal to the situation in

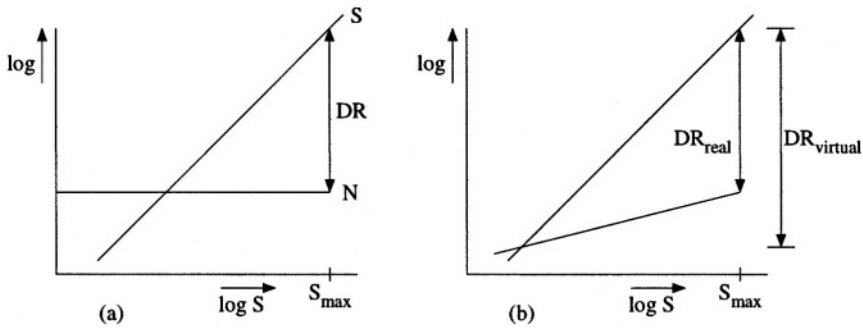

Figure 2.10: The influence of automatic gain control. a) The system without and b) the system with automatic gain control.

which no automatic gain control is used. For lower signal levels an improvement of the signal-to-noise ratio is obtained. The virtual maximal dynamic range is indicated by the most right arrow. It is the ratio of the largest and smallest signal which can be processed; however, *not* at the same time.

2.6.4 Bandwidth

The bandwidth of a circuit is limited by speed limitations of the devices used. Normally, the active devices limit the bandwidth of the circuit. However, in low-voltage low-power applications (i.e. supply voltage is 1 V and currents are on the order of μA and nA), resistors in the range of MΩ to GΩ become required. As these resistors are on-chip relatively large, the parasitic capacitance to the substrate makes these resistors frequency dependent with a relatively low bandwidth. As a consequence, resistors are also becoming a limiting factor for the bandwidth.

In contrast to the exchange of noise and signal power, it is not easy to exchange bandwidth for dynamic range (noise or signal power). This is because very often both the signal and noise power have to do with the signal amplitude whereas bandwidth has to do with speed of the signal; a totally different property. Such an exchange implies a rather large change at system level, as was already indicated. An example of bandwidth and dynamic range exchange is D-to-A and A-to-D conversion, see figure 2.11. Digital signals only have two signal levels. The noise may be relatively large. The bandwidth of digital signals can be very large, i.e. Gbit/s. Analog signals, however, are characterized by the fact that the signal level can have each value between a minimum and a maximum value. A relatively small noise signal can already cause a loss of information. The bandwidth of analog systems can generally be lower, but it has the same potentials as the digital signal concerning bandwidth.

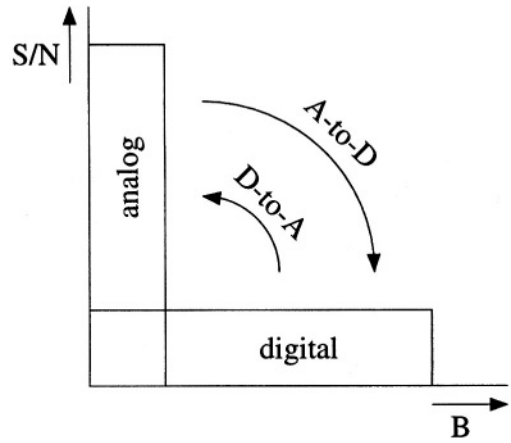

Figure 2.11: Digital and analog signals in the S/N-versus-B plane.

Changing the signals in a system from A to D or vice versa, has, as may be clear, a large influence on the total system.

2.7 Relation to the circuit's topology

In the previous sections the fundamental limits of the channel capacity were translated into:

- noise;
- distortion/signal power;
- bandwidth.

It is preferable to have these three quality aspects orthogonal in the process of circuit design. This practically means that each of these three aspects must be determined in parts of the circuit that do not completely overlap. It can be made plausible, with some simple reasoning, that this orthogonalization is possible, to a certain extent, of course.

Noise has the largest influence at those places where the signals are the smallest. Such a place always exists. For instance, in an amplifier this is mostly at the input. It is very often the place where the information source is found, or is to be connected to the circuit.

Clipping distortion is likely to be located at the stages where the signal excursions are the largest. Thus, the sensitive point for clipping distortion and noise are located at different places in the circuit. For the amplifier again, the stage determining the clipping distortion is very often the output stage.

Bandwidth concerns all the components in the circuit; whether a slow device is at the input or at the output, it is able to filter out the fast signals. Consequently, the parts not used to minimize noise and distortion can be used to optimize the bandwidth of the total system, this inherently assumes negative feedback, see also section 2.8.2.

From the previous discussion it follows that the noise and clipping behavior have to be designed first[4] before the bandwidth is maximized, because bandwidth can be realized without disturbing the noise and clipping behavior whereas designing the noise and clipping behavior without altering the bandwidth is not easy to do. It should be noted however, that when the noise behavior is designed, it does not mean that it may no longer be changed. For instance, at the cost of a small degradation of the noise level, a relatively large improvement of the bandwidth can be obtained, resulting in the saving of an additional stage.

2.8 Accurate circuit design

The design of the circuit starts with ideal blocks and the noise, distortion and bandwidth limitations are introduced in different design steps. At each level the models are refined and more non-idealities are introduced. To ease the design, only the relevant parameters should be used. For instance, when minimizing the noise level, it generally makes no sense to take the thermal noise of the collector bulk resistance into account.

2.8.1 Key parameters

The key parameters are those parameters whose values have to be known accurately as they predominantly determine the behavior. The other parameters only have to be large or small enough. For instance, for the design of bandgap references, V_{AR}, I_S, E_G and X_{TI} are the key parameters. The influence of all the other parameters can be made negligible.

The identification of the key parameters helps the designer to reduce the complexity of the design process as the number of parameters is reduced. If, subsequently, the circuit is realized such that only the key parameters that were previously found determine the behavior of the circuit, the spread in its behavior is also reduced. On top of this, the designer can more easily derive the optimum values for the key parameters. Then, these values can be the aim for the technological people who want to optimize their processes. This results in processes that are optimum for specific circuits. This is called design-driven technology.

[4] In [6] it is made clear that doing the noise minimization before minimizing the clipping distortion is the order to be preferred.

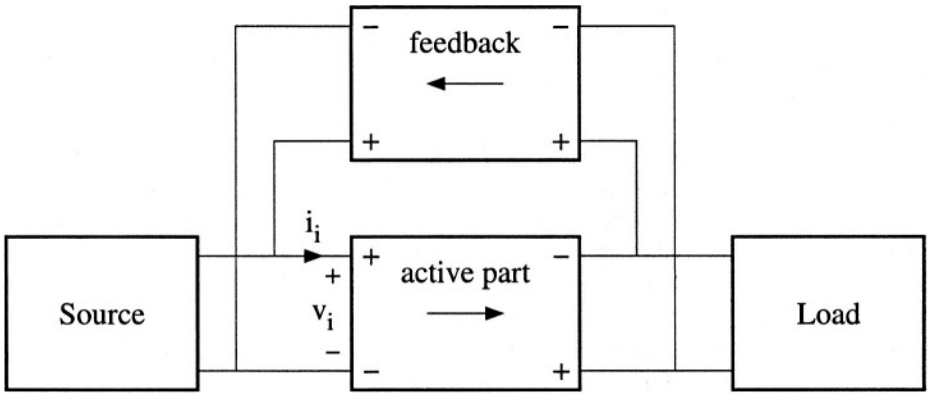

Figure 2.12: Negative feedback.

The number of key parameters has to be as low as possible, in which case the design is kept as simple as possible. At device and layout level the designer has the ability to reduce the number of key parameters at relatively low cost.

At device level, the operating point has to be chosen such that the smallest number of parameters influences the behavior of the device. For instance, when the operating point of a transistor is close to high-level injection, more key parameters are found. This is simply circumvented by choosing a lower biasing current or a larger transistor.

From the layout point of view, more specific measures can also be taken to reduce the number of key parameters. For instance, for a transistor in a low-noise application, using more base contacts in order to reduce the base resistance is a relatively simple way to reduce the noise level. At circuit level several devices could be taken in parallel. However, this would lower the speed of the transistor.

2.8.2 Negative feedback

The previous section detailed the motivation behind the use of key parameters for the design of circuits. This section deals with negative feedback, a synthesis method to reach the transfer described by the key parameters as closely as possible [7], [8] and [9] with relatively little effort and to realize more orthogonality in the design process. The principle of negative feedback is depicted in figure 2.12. A system with negative feedback consists, besides a source and a load, of a feedback network and an active part. The feedback network sets the ideal transfer and the active part provides for the required power gain. The ideal transfer is found when the active part is ideal, i.e. infinite power gain. The

nullor [10] is a network-theoretical element having this infinite power gain. It is a combination of a nullator and a norator. The nullator introduces an extra constraint, i.e. $v_i = 0$ and $i_i = 0$, whereas the norator provides for the additional degree of freedom in the circuit by its output voltage and current, v_o and i_o, respectively, to fulfill the extra constraint.

A very convenient way of describing amplifiers is by means of the chain matrix. It is an anti-causal description and defined as:

$$\begin{pmatrix} v_i \\ i_i \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \cdot \begin{pmatrix} v_o \\ i_o \end{pmatrix}, \tag{2.6}$$

with A, B, C and D the chain parameters. The advantage of using the chain parameters is that the chain matrix of a cascade of two blocks is simply found by multiplying the corresponding chain matrices.

For the nullor all the chain parameters are zero. The nullor can only be used in a sensible way when feedback is used. Otherwise, each input signal not equal to zero results in an infinite output signal. The transfer of the feedback system with the nullor as given in figure 2.12 equals:

$$\frac{E_{load}}{E_{source}} = \frac{1}{\beta(s)} = A_{t\infty}, \tag{2.7}$$

where $A_{t\infty}$ is the ideal transfer and $\beta(s)$ is the transfer of the feedback network. In the case of a non-ideal nullor implementation, the transfer is found to be:

$$\frac{E_{load}}{E_{source}} = A_{t\infty}\frac{-A(s)\beta(s)}{1 - A(s)\beta(s)} - \frac{\rho(s)}{1 - A(s)\beta(s)}, \tag{2.8}$$

where $A(s)$ is the transfer of the active part and $\rho(s)$ is the direct transfer from the input to the output. The product $A(s)\beta(s)$ is called the loop gain of the system. When this loop gain approaches infinity, the transfer of the system approaches $A_{t\infty}$. This term is therefore called the asymptotic gain.

When the gain of the active part is large, the transfer is set by the transfer of the feedback network. The parameters of this network are therefore key parameters for the transfer of the amplifier. The parameters of the active part are not important for the transfer function. For the quality of the transfer, i.e. noise, distortion and bandwidth, key parameters are found inside the active part, which describe the behavior of the non-ideal amplifier with respect to these three criteria.

In the case of amplifier design, the feedback network is mostly implemented by accurate resistors, if necessary they must be off-chip, or resistor ratios, which can be relatively accurate on-chip (passive components). The gain part is implemented usually by inaccurate devices providing for power gain (active devices).

Feedback can generally be used to obtain an accurate relation between the actual signal-processing function and the components intended to realize it. For

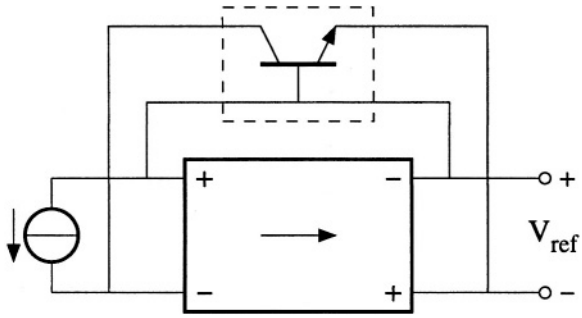

Figure 2.13: An accurate relation between collector current and base-emitter voltage by means of negative feedback.

instance, if the transfer from collector current to base-emitter voltage is desired for realizing a reference voltage, a system as given in figure 2.13 is obtained. The transfer between base-emitter voltage and collector current is used as the feedback transfer. When the gain of the active part is relatively high, the transfer from current to voltage is set by the transistor in the feedback network.

2.9 Homogeneous circuits

After the discussion about limitations, optimizations and design, again the systems of which the homogeneous solution is important are discussed and the corresponding basic electronic functions are derived. One common aspect of these basic electronic functions was already found: the blocks do not need an input signal.

The following sections discuss the general solutions of several orders of homogeneous linear differential equations, starting at order one. It shall be seen that at this level of hierarchy, solutions can already be ignored when the three fundamental criteria are considered. Some solutions show an inherently non-ideality with respect to, mainly, noise.

2.9.1 First-order differential equation

The lowest order of differential equations is the first-order one. The general expression for the homogeneous linear first-order differential equation is given by:

$$a\frac{de}{dt} + be = 0. \tag{2.9}$$

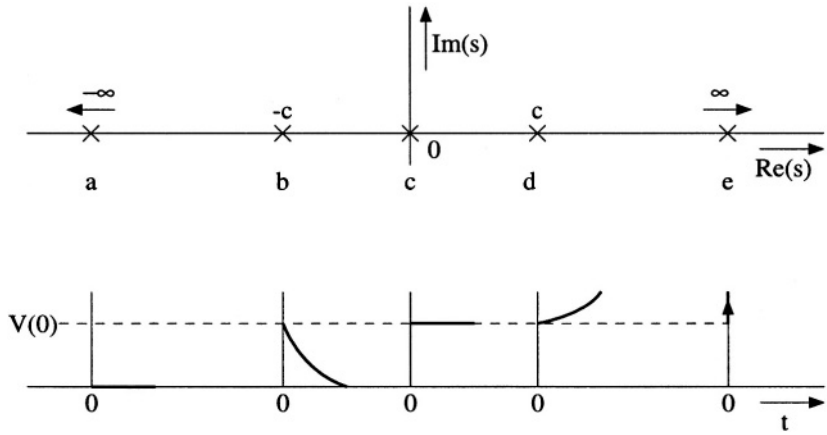

Figure 2.14: The solutions of the first-order differential equation in the Laplace domain and time domain.

The general solution is given by:

$$e(t) = A \exp\left(-\frac{b}{a}t\right), \tag{2.10}$$

where A is a constant. This differential equation has one differentiator. The output of a differentiator is the difference between the actual value and the previous value of e. With "previous" an infinitesimally small time in the past is meant. Thus, a differentiator introduces a memory function into the system. For this memory function the initial state has to be given in order to find the solution of the system described. When the initial condition is assumed to be $e(0)$, the solution is found to be:

$$e(t) = e(0) \exp\left(-\frac{b}{a}t\right). \tag{2.11}$$

Now all kinds of values of a and b have to be tried to find all the different types of solutions. When the differential equation is solved via the Laplace domain, the solutions are almost trivial. The Laplace transformation of the solution equals:

$$E(s) = \frac{E(0)}{1 + \frac{b}{a}s}. \tag{2.12}$$

From this expression five different types of solutions can be found. These solutions, with the corresponding time domain functions, are depicted figure 2.14. The five solutions are:

a) $s = -\infty$;

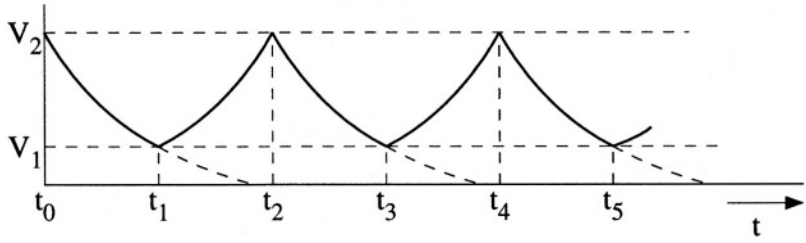

Figure 2.15: An oscillator realized by switching between two poles.

b) $s = -c \wedge c \neq \infty$;

c) $s = 0$;

d) $s = +c \wedge c \neq \infty$;

e) $s = +\infty$.

Two of the five solutions can be ignored immediately. The pole at minus infinity results in an output signal that is always zero. It describes a system with an infinite bandwidth and no input. The output of the system goes instantly from the initial state to zero. Analogously, the output of the system with the pole at plus infinity goes instantly to infinity.

Solutions b and d result in an exponentially decreasing and increasing signal as a function of time. The initiation is caused by the initial condition of the system. When nothing additional is done, these signals are of no use. However, when two detection levels and a sign switch are introduced as indicated in figure 2.15, an oscillator can be made. During t_0 to t_1 the output signal decreases exponentially to zero. However, when the output signal reaches the detection level V_1, the sign of either a or b from the differential equation is changed. The output signal now increases exponentially from V_1 to V_2 and again a or b is changed in sign. This results in a periodic signal whose frequency is determined by the poles and the detection levels. The system flips between situations b and d of figure 2.14. This oscillator has two principal problems:

- the oscillator only starts when the output signal at t_0 is larger than V_1. This can be circumvented by using some kind of predefined initial condition;

- the oscillator is characterized by a pole, flipping from the right half plane to the left half plane, with a real part not equal to zero. This real part can only be realized by dissipation. Dissipation is only possible in real transfers, i.e. resistive. Thus, in principle, this oscillator *cannot* be noise free. Its noise level can only be made low at the cost of power.

Especially because of this last item, this frequency constant is not discussed any further.

It should also be noted that this oscillator is, in principle, described by a non-linear differential equation. For the sign switching a memory function is required, which would obviously increase the order of the differential equation to two. However, as the frequency behavior of this oscillator is determined by one pole at a time, it can be called a first-order oscillator.

Solution c is the single pole at the origin. It is an ideal integrator. As no input signal is present, the output signal is constant and equal to the initial condition. The system memorizes the voltage present at $t = 0$. When the initial condition is given a reference value, solution c acts as a constant reference. At this level, the reference can be noise free as no dissipation is involved.

Summarizing, from the linear homogeneous first-order differential equation, the only ideal system found is the constant reference. It should be noted that the first-order oscillator as described by [11] does not belong to the solutions presented in this section. This oscillator is based on one pole at the origin and a constant *input signal,* the integration constant. For obtaining a frequency reference the sign of the integration constant is changed as a function of the output signal of the integrator. This oscillator is described by the *non-linear* van der Pol equation [12]. As the integration constant, i.e. the input signal of the integrator, depends on the output signal of the integrator, ánd the integrator has either the dimension $j\Omega$ or jS, for the capacitor and inductor, respectively, the transfer from the output to the input of the integrator by means of the sign switch and integration constant, has a resistive part. Reactive elements cannot be used here as they would increase the order of the system. Thus this oscillator cannot be noise free, fundamentally.

2.9.2 Second-order differential equation

When a system consists of two differentiators or integrators, a second-order differential equation is found describing its behavior. The general homogeneous linear second-order differential equation is given by:

$$a\frac{d^2e}{dt^2} + b\frac{de}{dt} + ce = 0. \tag{2.13}$$

This differential equation has a maximum of two different roots. These roots can be real or complex conjugated. The solution based on real roots are also ignored here as they fundamentally introduce noise in a system. The solution with the complex conjugated poles, $s = \sigma \pm j\omega t$, is given by:

$$e(t) = \exp(\sigma t) \cdot [A\cos(\omega t) + B\sin(\omega t)]. \tag{2.14}$$

This solution describes a harmonic signal whose amplitude as a function of time is determined by the term $\exp(\sigma t)$. For $\sigma < 0$, the amplitude decreases as a

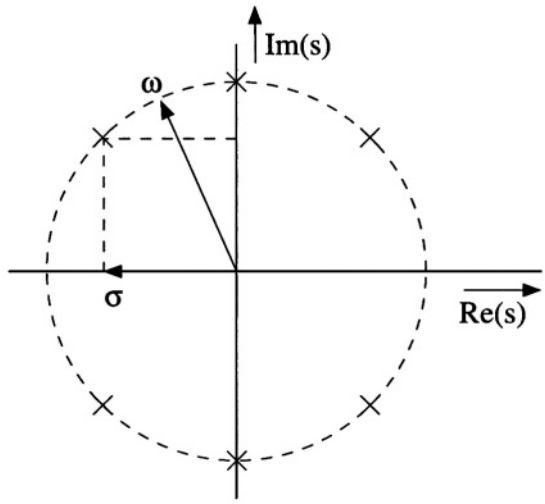

Figure 2.16: The poles in the s-plane for the three different solutions having complex conjugated poles.

function of time due to dissipation. Consequently, again, this solution cannot be noise free. For $\sigma > 0$, the amplitude increases as a function of time and this is a non-practical solution. Finally, the solution with $\sigma = 0$ is a harmonic signal with a constant amplitude as for $\sigma = 0$ the exponential term becomes one. When the two complex conjugated poles are both at the origin, the system also becomes a constant reference source. These three situations are depicted in figure 2.16.

The second-order differential equation has two differentiators and consequently two initial conditions are required to characterize the complete system. The initial conditions determine in the case of the harmonic signal the amplitude and the phase at $t = 0$, for instance. For the constant reference, the two initial conditions determine the value of the constant output signal.

Summarizing, the second-order homogeneous linear differential equation has one ideal solution, the harmonic frequency reference.

2.9.3 Third and higher-order differential equation

From the two previous orders of differential equations, it may have become clear that only those systems in which all the poles are on the imaginary axes can fundamentally be noise free. For a noise-free third-order system one pole must always be at the origin and the two complex conjugated poles must be somewhere else on the imaginary axis, see figure 2.17. These solutions do not

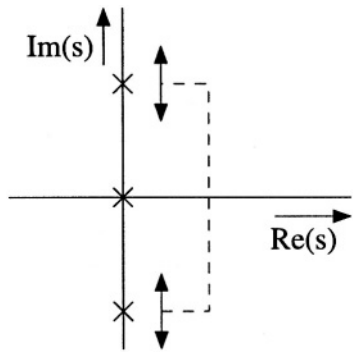

Figure 2.17: The poles of a noise-free third-order system.

introduce any new functions. When all the poles are at the origin, again, a constant reference is found. When one pole is at the origin and the other two poles are on the imaginary axis, the result is a harmonic frequency reference. The pole at the origin only introduces a 90° phase shift between the internal signal and the output signal.

For the fourth and higher-order differential equations, the discussion is analogous to the discussion for the third-order differential equations. No new functions are found, other than the constant and frequency references.

2.10 The amplifier

From the homogeneous differential equations two function blocks were found:

- the constant reference;
- the harmonic frequency reference.

When implementing these functions, it will be seen, as discussed in the corresponding chapter, that the amplifier is an inevitable function. For instance, it was seen that the constant reference can be considered as the memorizing of an initial condition. When this reference is loaded and no appropriate buffering is done, the reference quantity may change due to the dissipation of the load. This buffering requires an amplifier function with transfer one.

For the harmonic frequency reference two poles need to be exactly on the imaginary axis. This is only possible when absolutely no dissipation is present in the circuit. However, in practice some dissipation is unavoidable. To prevent the output signal from decreasing in amplitude, the losses need to be compensated for. The power losses are a function of the momentary amplitude and thus the

additional energy supplied should also be a function of the amplitude. For this, an amplifier is required [13].

Therefore, the amplifier will be discussed in chapter 4 before the harmonic frequency and constant references are discussed.

2.11 Conclusion

To be able to design high-performance electronic circuits a structured design method becomes inevitable nowadays. Structured electronic design requires a language that describes the required signal-processing function independently of (objective and unambiguous) parameters which describe the quality of this processing. This chapter showed the mathematical description language using differential equations to be the appropriate language. In this way it is possible to introduce the different quality aspects into the design process one by one. Moreover, the complete mathematical tool box becomes available for the designer in his struggle to find a circuit fulfilling all the requirements.

Subsequently, electronic limitations were introduced. These can be divided into three types:

- design requirements. Requirements which the electronic solution has to fulfill at least. These requirements are derived from the application;
- practical limits. Limits imposed by technological aspects, for instance. These limits can be relaxed by choosing a different process for integration, for instance;
- fundamental limits. These limits set the utmost performance a system can ever reach. All possible limits can be gathered under the three fundamental limits as given by Shannon: noise, signal power and bandwidth. These three limit the signal-transporting capability of a system, fundamentally.

These three types of limitations set the space of solutions in which the designer has to seek his solution. In this chapter the search method by creation was shown to fit the mathematical language the best in order to find the optimum solution mathematically.

Inherently to structured design is orthogonalization. It ensures that the three criteria, noise, signal power and bandwidth can be optimized separately.

The last part of the chapter derived the functional blocks that correspond to the solutions of homogeneous linear differential equations. The following were found:

- constant reference (from the first-order differential equation);
- harmonic frequency reference (from the second-order differential equation).

Other solutions are ignored, mainly on noise considerations, as they already show non-ideal behavior at the high level as discussed in this chapter. Finally, it was argued that an important and complicated building block for realizing both homogeneous functional blocks is:

- the amplifier.

Bibliography

[1] M.H.L. Kouwenhoven, G.L.E. Monna, C.J.M. Verhoeven, and A.H.M. van Roermund. Design of optimal dynamic range integrated mixer-filter systems. In *Proceedings of the IEEE International Symposium on Circuits and Systems,* volume 5, pages 261–264, 30 May – 2 June 1994.

[2] A.H.M. van Roermund. Chapter 1, Introduction. In W.A. Serdijn, C.J.M. Verhoeven, and A.H.M. van Roermund, editors, *Analog IC Techniques for Low-Voltage Low-Power Electronics.* Delft University Press, 1995.

[3] A.P. Chandrakasan and R.W. Broderson. *Low Power Digital CMOS Design.* Kluwer academic publishers, Boston, 1995.

[4] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal,* 27(3):379–432 and 623–656, July 1948.

[5] G. Groenewold. *Optimal Dynamic Range Integrated Continuous-Time Filters.* PhD thesis, Delft University of Technology, March 1992.

[6] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

[7] H.S. Black. Stabilized feedback amplifiers. *The Bell System Technical Journal,* pages 1–18, January 1934.

[8] H.S. Black. U.S. Patent 2, 102,761, 1937.

[9] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[10] H.K. Carlin. Singular network elements. *IEEE Transactions on Circuit Theory,* 11:67–72, March 1964.

[11] C.J.M. Verhoeven. *First-Order Oscillators.* PhD thesis, Delft University of Technology, February 1990.

[12] B. van der Pol. The nonlinear theory of electric oscillations. *Proceedings of the Institute of Radio Engineers,* 22(9):1051–1086, September 1934.

[13] C.A.M. Boon. *Design of High-Performance Negative-Feedback Oscillators.* PhD thesis, Delft University of Technology, September 1989.

Chapter 3

Low-voltage low-power design

3.1 Introduction

For many years designers have been interested in the field of low-voltage low-power design [1], [2], [3] and [4]. Very complex electronic systems are becoming smaller and even portable. The hand-held phones used nowadays, for instance, are very small. To be able to realize such small systems, a dominating volume-consuming component, the battery pack, has to be reduced in size. This size reduction can be done in two ways. Firstly, the power consumption of the electronic circuits can be lowered so that the *size* of the individual batteries reduces and, secondly, the supply voltage of the circuits can be lowered so that the *number* of batteries reduces. Battery voltages are on the order of 1 V - 1.5 V and, as one battery is the smallest number of batteries to supply the power for a circuit, electronic design is focusing on 1 V design combined with low power consumption. When a new type of battery is to be developed, having a higher voltage for equal or less volume, the influence of a low-voltage constraint is reduced to a large extent.

From standard digital design additional drives are found for lowering the supply voltage. The power consumption of digital circuits is proportional to the square of the supply voltage and thus reducing the supply voltages is a very effective method for reducing the power consumption. Further, as a result of the down scaling of devices, device operating voltages which, for digital circuits, very often equal the supply voltage, have to reduce. Breakdown voltages scale, approximately, linearly with the scaling of the device sizes. For modern sub-micron MOS-transistors, the breakdown voltages are on the order of only a few volt and supply voltages have to be well below them. If analog and digital

circuits are to be integrated on the same chip, and only one power supply is to be used, the supply voltage of the analog part has to reduce too.

In this chapter, the influence of a low-voltage, low-power constraint on the circuit's performance is examined. Low-voltage and low-power design are not completely independent topics. Low-power design may also concern low-voltage design, only the additional constraint low current is introduced. Therefore, a better distinction is:

- low-voltage design;
- low-current design.

To get a more clear (orthogonal) insight into the influence of a low-voltage, low-power constraint, these two issues, *low voltage* and *low current,* are discussed before *low power.* It will be shown that low-voltage design mainly alters the topology of the circuits, whereas a low-current constraint mainly lowers the performance with respect to noise, distortion and bandwidth.

3.2 Physical information carriers

In an electronic circuit the information can be coded in four different domains. These domains are:

- voltage;
- current;
- charge;
- flux.

Charge, q, is fundamentally related to current, i, via:

$$i = \frac{dq}{dt}, \tag{3.1}$$

and flux, ϕ, is fundamentally related to voltage, v, via:

$$v = -\frac{d\phi}{dt}. \tag{3.2}$$

Thus current and voltage can be said to be the prime domains. When information is coded on either charge or flux and this information is transported, time-varying charges and fluxes are found. This inherently means that currents and voltages arise.

To be able to do signal processing with electronic devices, power has to be supplied to them. Two types of power supplies can be distinguished:

- voltage supply → free current;
- current supply → free voltage.

These power supplies have one free variable to be able to supply a continuous range of power levels. These are the current and voltage for the voltage and current supply, respectively. The voltage and current domain of the voltage and current source, respectively, are bounded as they set a maximum on the magnitude of the voltage and current, respectively. Therefore, using the domain of the free variable as the domain for the information coding, more freedom and larger dynamic ranges can be obtained compared with the freedom and the dynamic ranges that would be obtained when using the bounded domain. In the case of a linear relation between the voltages and currents, the corresponding dynamic ranges in the voltage and current domain are equal. In the case of a specified dynamic range for the free domain, the impedance *level* can be used to make the relation between the voltages and currents such that the corresponding dynamic range also fits in the bounded domain. For example, in the case of a supply *voltage,* the currents can be chosen freely. The dynamic range of the voltage and current representations of the information are equal but the signal voltages are directly limited by the supply voltage, whereas the maximum current can be increased to infinity, fundamentally, by altering the impedance level.

For a non-linear relation between voltages and currents, the corresponding dynamic ranges differ; the dynamic range in the bounded domain can be smaller or larger, i.e. compression or expansion, respectively, compared with the dynamic range in the unbounded domain (it is assumed that the unbounded domain is used as the domain for the information coding). As non-linear relations are involved, the term dynamic range can no longer be used unconditionally (cf. paragraph 2.6.3).

The commonly used electronic devices have an expanding function from input voltage to output current. Supply sources are mostly of the voltage source type: current is the free domain. These are probably the two dominant reasons why low-voltage current-mode circuits [5] have become very popular during the last few years. Current-mode design *fits* to electronic design with power-supply voltages and the commonly available electronic devices. Therefore, it is the way to reach a large dynamic range together with the low-voltage constraint.

Finally, it should be noted that low-voltage design does not inherently mean current-mode design. For a system whose input signal is a voltage, the processing should preferably be done in the voltage domain, unless there are reasons as mentioned before, for processing in the current domain. When going from the voltage to the current domain or vice versa, using transformations comprising a resistive part, extra noise is introduced. Therefore, unnecessary transformations should be avoided to prevent the system from sub-optimal performance.

For the remainder of this chapter a power-supply *voltage* is assumed.

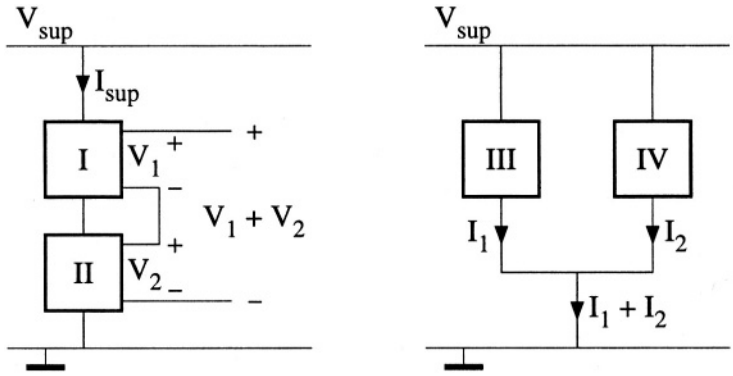

Figure 3.1: The change from series-connected blocks to parallel-connected blocks for the addition of two signals for low-voltage design.

3.3 Low voltage

In the history of electronic design, low-voltage design may be seen as a separate period [1]. After the introduction of IC technology in electronic design, circuits changed with respect to circuits previously realized with discrete components. For integrated circuits, transformers, inductors larger than some tens of nH and capacitor values beyond several hundreds of pF are a problem. These components cannot be integrated and have to be mounted externally to the chip. Consequently, when fully integrated circuits have to be realized, components with such values can no longer be used. In contrast to this *reduction* of design freedom, the matching of components on chip is far better than the matching of discrete components and has led to a profitable use of this feature. These aspects caused the circuit's topology, appearance, to change.

Something analogous is now happening after the introduction of low-voltage design; the appearance of circuits is changing. Circuit solutions that previously were realized with a number of transistors stacked are no longer possible due to the lowered supply voltage. New circuit topologies are found that realize the same functionality, but now with more parallel current paths, see figure 3.1. Instead of a common supply current, I_{SUP}, blocks have a common supply voltage, V_{SUP}.

[1] The circuits in which tubes are used as the active devices can also be seen as low-voltage circuits. For these circuits the stacking of devices is not preferable as it results in very high supply voltages. Different techniques are used, compared with the modern low-voltage design, to cope with this.

3.3.1 Implications at signal-processing level

Generally speaking, the low-voltage constraint sets a limit on the maximum magnitude of all the node voltages. Finally, these node voltages can be a result of a signal and/or a required quiescent point. At the signal-processing level the quiescent point of devices and larger blocks are not present yet and, consequently, only the signals in the system can conflict with the low-voltage constraint.

As was discussed in the previous section, when the current is used as the signal carrier, the low supply voltage is no longer a direct limitation. However, it is not always possible to do the complete signal processing in the current domain. For instance, if the information of a sensor is best retrieved when its output *voltage* is used, voltage swings cannot be made arbitrarily small. They are at least equal to the sensor signal voltages. The same holds for an output transducer which reproduces the information the best when a voltage is supplied, the supply voltage must be at least as high as the highest output voltage.

Between the input and output transducer, voltage swings can be minimized in a mathematical way. The input signals are known, the signal processing function is known and thus the internal voltages can be found (it is assumed that the complete differential equation is already reduced to a set of lower order differential equations). By means of the associative and the commutative laws, for instance, internal voltages can be minimized.

3.3.2 Implications at circuit level

At the signal-processing level only the trivial influence was found that node voltages are limited to the supply voltage. Concerning the influence on the circuit's performance more can be said. As was already mentioned, the influence of the low-voltage constraint is mainly found in the changed topology of the circuits. Due to the low-voltage constraint, combinations of transistors previously used can no longer be used. This, however, does not necessarily have to reduce the circuit's performance, for instance, implementing nullors by cascading CE-stages [6] remains possible.

A circuit can be separated into its signal path, the components primarily required for the implementation of the signal-processing function, and the remaining part of the circuit which is required to make the signal path function properly, the so-called bias circuit. The supply voltage now has to be divided between the circuit for the signal path and for the bias circuit, see figure 3.2. In the figure, v_{signal} stands for the voltage required for the complete signal range. The information may be directly related to this voltage or via an impedance. V_{DEVICE} stands for the voltage required for the devices of the signal path in order to guarantee the required small-signal behavior (this is discussed in the next section). The voltage margins at the top and the bottom are the margins

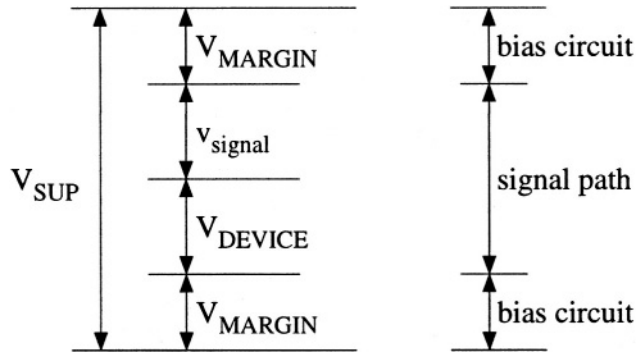

Figure 3.2: The division of the available supply voltage between the signal-path and bias circuit.

required for proper functioning of the bias circuits. The value of these margins depends, amongst other things, on the type of biasing used. For instance, for class A and AB biasing these values are determined by the saturation voltages of the current sources. Whereas for class B biasing the bias sources are no longer present and V_{MARGIN} can therefore be set to zero. Instead of these sources, signal path-devices are now also found. Either one of the two devices is conducting and the other is switched off.

For a voltage-mode circuit the signal-voltage swing is directly related to the dynamic range and, consequently, the supply voltage should be used as much as possible for this signal-voltage swing, otherwise a waste of power results [7]. This maximization of the voltage swing is limited by the minimum voltage margins required for the bias circuits and the minimum voltage for the devices of the signal path. Increasing the voltage swing so that these margins or the voltages for the signal-path devices are no longer available results in a power increase required for getting them properly functioning again, if this is still possible. Consequently, a reduction of the power efficiency is obtained.

For a current-mode circuit the power used for the signals, the signal-path devices and the bias circuits can be depicted as illustrated in figure 3.3. Figure 3.3a depicts the ideal situation. This means that the signal-path devices and the bias circuits also function properly at zero voltage and the impedance level can be made zero, such that the voltage swing becomes zero, without the cost of extra power. In that case the total power consumption is constant and equals $i_{signal} \times V_{SUP}$. However, in practical cases, the signal-path devices and the bias circuits require some minimal voltage for proper functioning. When this minimal voltage is not available, (very) much additional power is required (including additional circuits) to get the circuit functioning, if possible. For obtaining very small voltage swings a relatively low impedance level is required. This also

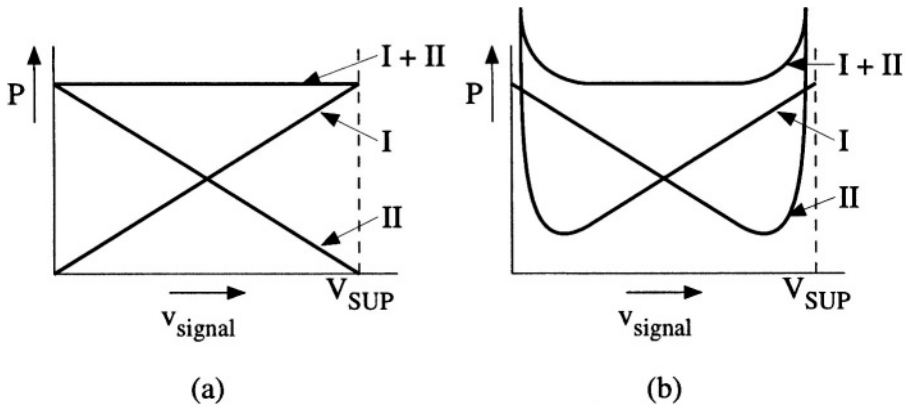

Figure 3.3: Schematic representation of the power consumption for a given signal current: I, the power used for the signal and II, the power used for the signal-path devices plus the bias circuit. a) In the ideal case and b) the non-ideal case.

costs additional power. Therefore the total power consumption (the efficiency) reduces for relatively small and large voltage swings.

Thus for both the voltage-mode and current-mode circuits it is preferable to have minimal V_{MARGIN} and V_{DEVICE}. For the voltage-mode circuits, the dynamic range can be increased, and for the current-mode circuits the supply voltage can be reduced, resulting in less power consumption.

The functions to be discussed in this book can all, besides the principal elements for generating the reference variable, be implemented by a combination of nullors and passive linear elements. At the signal-processing level only the signal levels could result in a conflict with the low-voltage constraint. For practical realizations of this signal path with realistic nullor implementations, a minimum voltage is required to be able to work correctly, i.e. V_{DEVICE}. Consequently, conflicts may arise between the node voltages and the low-voltage constraint. The linear passive elements do not require a supply voltage [8], they already show the required transfer at a zero voltage.

Thus, to gain insight into the influences of a low-voltage constraint on the circuit's performance, the influence on the nullor implementation needs to be examined and the influence on the performance of the bias sources has to be known. Then an optimization can be done on the nullor implementation combined with the bias circuit.

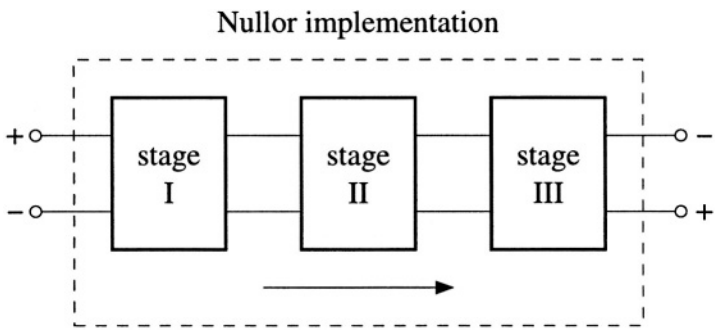

Figure 3.4: A nullor implemented by a cascade of stages.

3.3.2.1 The nullor implementations

The required voltage at the input and output of the nullor implementation, V_{DEVICE}, and the voltage margins for the bias circuits, V_{MARGIN}, determine, together with the actual supply voltage, the maximum voltage swing, v_{signal}, at the input and output, respectively. As was pointed out, minimizing V_{DEVICE} and V_{MARGIN} is favorable for voltage-mode as well as current-mode processing. Therefore, the focus can be on the minimum supply voltage required for a nullor implementation. The difference between this minimum supply voltage required and the actual supply voltage can be used for signal processing.

Nullors can be considered to be comprised of a number of cascaded stages, see figure 3.4. Each stage in turn is realized by means of one or more devices. To obtain the optimum contribution of a stage to the performance of the total nullor implementation, it should be without local feedback [6]; the four chain parameters should be as small as possible. The only stages having this feature are the CE stage for the BJTs and the CS stage for the (MOS)FETs. The CB, CG and CC, CD stages are CE and CS stages with a (non-energetic) local feedback so that they act as a current follower and a voltage follower, respectively.

The devices used for the CE and CS stage require a bias in order to obtain the desired small-signal performance [8]. In figure 3.5 the required bias is depicted for the BJT and the (MOS)FET. For a bipolar transistor this bias is given by: the base-emitter voltage V_{BE}, the base current I_B, a minimal collector-emitter voltage V_{CE} and the collector current I_C. For the MOS(FET) this bias is given by the gate-source voltage V_{GS}, the gate current I_G, a minimal drain-source voltage V_{DS} and the drain current I_D. For the bipolar transistor the four parameters are free to choose with the constraint that two degrees of freedom are available [9]. For the (MOS)FET the gate current is a leakage current and is not free to choose. Therefore, for the (MOS)FET three parameters are free to choose but also with the constraint of two degrees of freedom.

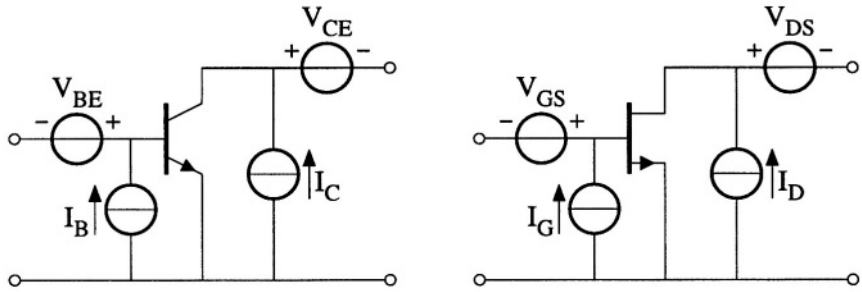

Figure 3.5: The bias of a BJT and a (MOS)FET.

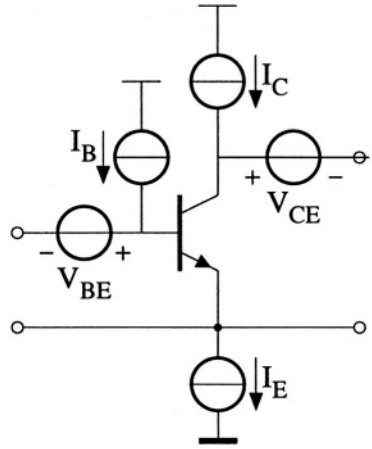

Figure 3.6: A more practical configuration for realizing the bias of a bipolar transistor.

As the collector/drain current is most closely related to the behavior of the device (see section 3.4) and the collector-emitter/drain-source voltage is important for the large signal behavior (preventing saturation) it is the most convenient to set these two quantities and let the other(s) be controlled by a loop [8]. Figure 3.6 gives a more practical configuration for realizing the bias (it is depicted for the BJT only). Instead of realizing floating current sources, each current source is split into two sources, each having one terminal grounded.

For this stage the minimum supply voltage required at the input is found to be equal to:

$$V_{SUP_{minimal}} = V_{DEVICE} + V_{MARGIN} = V_{BE} + V_{SAT_B} + V_{SAT_E}, \qquad (3.3)$$

where V_{SAT_B} and V_{SAT_E} are the saturation voltages of the base-current source and the emitter-current source, respectively. At the output the minimum supply

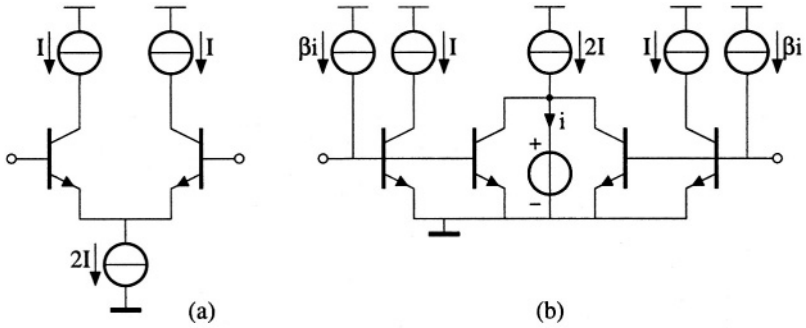

Figure 3.7: Realizing a balanced CE stage. a) In the voltage domain and b) in the current domain.

voltage required equals:

$$V_{SUP_{minimal}} = V_{DEVICE} + V_{MARGIN} = V_{CE} + V_{SAT_C} + V_{SAT_E}, \qquad (3.4)$$

where V_{SAT_C} is the saturation voltage of the collector-current source. The largest of the two determines the minimum supply voltage required for this stage. In a lot of situations the emitter of the CE stage is grounded and the current source I_E is short circuited and does not need to be realized; consequently, $V_{SAT_E} = 0$ V. Thus a single floating CE stage requires V_{SAT_E} more supply voltage than the grounded version.

Besides the single CE stage, the balanced version of the CE stage also has small chain parameters. The balancing can be done in two domains [10]:

- voltage domain;
- current domain.

Both these methods are depicted in figure 3.7. Figure 3.7a depicts the balancing in the voltage domain, it is the conventional differential pair. The balancing is a result of the anti-series connection of the inputs (voltage feedback) and the anti-series connection of the outputs of the two transistors. The sum of the two output currents is constant and equal to the tail current. For both transistors the biasing sources as depicted in figure 3.6 are required and thus the minimum supply voltages for the input and output are given by equations (3.3) and (3.4), respectively. Lowering this voltage by grounding the emitters, as was discussed for the single CE stage, is not possible as the balancing then disappears; the series feedback is broken.

For the method as depicted in figure 3.7b it is permissible to have grounded emitters. The two outer transistors are the signal transistors while the two inner transistors determine the sum of the currents through the signal transistors. As

a function of the resulting error, a *current* is fed back so that the sum of the currents of the two signal transistors becomes equal to 2I. Again, balancing is obtained, but now with current feedback. Circuits using this technique are for example [11] and [12]. For this stage the minimum supply voltages for the input and output are also given by equations (3.3) and (3.4), respectively. For this stage, as said, it is permissible to ground the common-emitter node, because the balancing is realized by means of the parallel feedback. Thus this balanced CE stage can have the same minimum supply voltage as the single CE stage. In appendix A, the minimum supply voltage required for the four different types of negative-feedback amplifiers is discussed.

3.3.2.2 The voltage required for the bias sources

Bias sources also require a minimum voltage to function properly. The bias sources can be a current source for, for instance, the bias current of the signal transistor or a voltage source used as, for instance, a level shift between two stages. In this section the relation between the supply voltage and the performance of current and voltage sources is examined at a relatively high level of hierarchy.

Voltage sources Bias voltages can be derived from the available voltage references, i.e. PTAT voltages, bandgap-reference voltages, supply voltages, et cetera. This derivation has no resistive part, i.e. from a voltage to a voltage, and thus the noise performance of the bias-voltage source can be as good as the voltage reference. In realistic sources, resistors, diodes and so on, are involved to realize the DC voltage conversion and introduce noise. The impedance level, however, can be chosen freely so that the noise level can be made as low as required.

Assume a bias voltage equal to V_{BIAS} has to be realized. This can be done by amplifying $(\mathbf{gain} \geq \mathbf{1})$ a reference voltage to the required value. When the supply voltage is now lowered and it is assumed that the required bias voltage remains constant and below the supply voltage, its performance is not influenced, i.e. the bias voltage is independent of the supply voltage. When the bias voltage is derived from the supply voltage by means of a divider, it becomes dependent on the supply voltage. An example of this type of source is depicted in figure 3.8. When the supply voltage is lowered, the division factor has to be changed by the same factor in order to get the original bias voltage again. Assuming an ideal supply voltage, the noise of this source is a result of the resistive divider. Its noise-power density spectrum, S_v, equals:

$$S_v = 4kT\frac{R_1 R_2}{R_1 + R_2}, \tag{3.5}$$

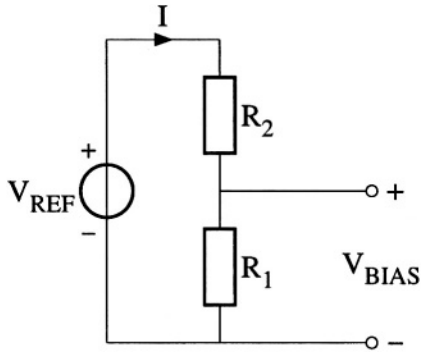

Figure 3.8: The realization of a bias voltage by dividing a larger voltage.

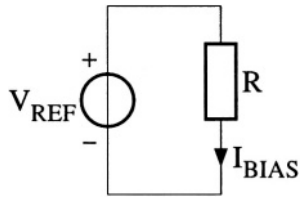

Figure 3.9: The derivation of a bias current from a reference voltage.

that is, for the noise the two resistors are in parallel. For supply voltages, relatively large with respect to the required bias voltage, R_2 can be ignored with respect to R_1 and the noise is constant as the current drain and the bias voltage are assumed to be constant. For supply voltages relatively close to the bias voltage, R_2 becomes relatively small and thus reduces the noise level. In the limit when the supply voltage equals the bias voltage, the noise level tends to zero. Thus the closer the bias voltage is to the supply voltage, the lower the noise level is. In other words, when the signal-to-noise ratio of the source is kept constant, the current consumption can reduce when lowering the supply voltage.

Current sources In nature no true current references are found and therefore bias currents have to be derived from voltage references, inherently introducing a resistive transfer, see figure 3.9. The reference source is again assumed to be ideal as it was in the case of the voltage bias source. The noise-power density spectrum, S_i, at the output of the current source equals:

$$S_i = 4kT\frac{I_{BIAS}}{V_{REF}}. \tag{3.6}$$

To obtain low-noise performance the reference voltage must be as large as possible and is thus a maximum when it equals the supply voltage. Thus for a given bias current the noise power increases proportionally with a decreasing supply voltage. By means of an increased power consumption, the equivalent noise of the resistor of figure 3.9 can be made lower.

3.3.3 Implications at device level

The performance of the individual devices is only slightly dependent on the supply voltage. In section 3.4 will be shown that the current through a device predominantly determines its behavior. As was pointed out in the introduction, lowering the supply voltage makes it possible to use smaller devices. As the maximum voltage in a circuit lowers, also the breakdown voltages of devices can be proportionally lower. Smaller devices can have smaller parasitic capacitances and have, consequently, higher speed potentials.

This increase in speed is slightly reduced by the increase of the capacitances of the junction which are normally reverse biased; this applies to bipolar transistors as well as to (MOS)FETs. For the capacitance of a reverse-biased junction, it holds [13] that:

$$C_j = C_{J0} \frac{1}{(1 - V_{EXT}/V_J)^M}, \tag{3.7}$$

where C_{J0} is the capacitance at zero voltage, V_{EXT} is the external voltage across the capacitor, V_J is the build-in voltage and M is the grading factor which depends on the doping profile. Values for M are about 0.3 and for V_J about 0.7 V. For example, for the DIMES01 transistor [14], the parameters for the base-collector capacitor of the minimal NPN transistor are given by $M = 0.28$ and $V_J = 0.67$ V. Expression (3.7) is plotted in figure 3.10 with $C_{J0} = 1$ pF. For a reverse bias voltage of about 8 V, the junction capacitance is halved.

The effective transit frequency $f_{T_{eff}}$ of a bipolar transistor is given by:

$$f_{T_{eff}} = \frac{g_m}{2\pi(g_m T_F + C_{je} + C_{jc} + C_{js})}, \tag{3.8}$$

where g_m is the transconductance of the transistor, T_F is the base transition time and C_{je}, C_{jc} and C_{js} are the junction capacitances of the base-emitter, base-collector and base-substrate junction, respectively. For vertical transistors, the capacitor to the substrate is found at the collector. When calculating the f_T of a transistor, the substrate capacitor is short circuited (when the collector bulk resistor is ignored), and does not need to be taken into account. But when the transistor is driven by another vertical transistor, the substrate capacitor of that transistor is in parallel with the base-emitter capacitance and as a consequence lowers the speed of the transistor. For the lateral transistors the capacitance to the substrate is already at the base and therefore equation (3.8) is used for both vertical and lateral transistors.

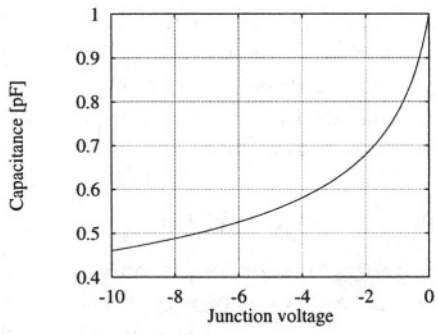

Figure 3.10: The capacitance of a reverse-biased junction as a function of the bias voltage, with $M = 0.28$, $V_J = 0.67$ V and $C_{J0} = 1$ pF.

For a lower supply voltage, the voltage that can be used to reverse bias the collector-base capacitor and the capacitor to the substrate also reduces, resulting in a lowering of the $f_{T_{eff}}$. The relative lowering depends on the value of the base-emitter diffusion capacitance, $g_m T_F$, which increases for increasing collector current, and on the value of the base-emitter junction capacitance, which increases for increasing transistor sizes (this also holds for C_{jc} and C_{js}).

3.4 Low current

In contrast to a low supply voltage, a low current consumption has a large influence on the performance of circuits. The performance of the individual devices is mainly determined by the values of their bias currents and degrade for lower currents. Before this is treated in more detail, first the implications at system and circuit level are discussed.

3.4.1 Implications at system level

For a power supply *voltage,* the current drain is a free variable. It is not in principle limited to a certain boundary value. Thus a low current consumption has no influence on the system level. If a circuit were to be supplied by a current source, then the current consumption of the circuit has to be fixed which is less trivial constraint than a fixed supply voltage.

3.4.2 Implications at circuit level

At circuit level, more or less the same applies as for the signal-processing level. But at this level, circuit techniques can be used to maximize the efficiency of

the current consumption. Straightforward efficiency improvement is obtained when instead of class A biasing, class (A)B biasing is used.

With class A biasing, the currents supplied to the signal path at least equal the level required for the maximum signal levels. Therefore, when relatively small signals are processed, a lot of current is wasted. For class (A)B biasing, this bias level can be said to be made signal dependent. When no signals are processed, the bias drops to zero (class B) or to a relatively small stand-by level (class AB). In this way the waste of current is considerably reduced.

In appendix B a design example is given of a balanced three-stage output amplifier whose last two stages are class AB biased [15]. The quiescent current of these stages is about 1% of their maximum output current. For realizing the AB relation additional circuitry is required, consuming current so that the total quiescent current amounts to about 5% of the maximum output current. But this is still an improvement of about a factor of 20 over the same amplifier with class A biasing.

3.4.3 Implications at device level

The bias current of a device determines its noise level, its speed and its output power. Consequently, a reduction of the current consumption can result in a tremendous reduction of the information handling capability. In the following sections these three design aspects will be discussed separately.

3.4.3.1 Noise

Reducing the current level of a circuit inherently results in the lowering of the signal-to-noise ratio of the signals. This can very easily be seen when it is remembered that the signal power reduces quadratically for reducing current level, whereas the noise power reduces linearly for reducing current level, see equation (3.9):

$$\frac{S}{N} \propto \frac{I^2}{I} \propto I. \tag{3.9}$$

Thus the signal-to-noise ratio decreases in proportion to a decreasing current level.

The resistor Validation of equation (3.9) for the case of a resistor is very easily done. By examining the resistor, an impedance level is inherently introduced and both the voltage and current may serve as the information carrying quantity. When the voltage is used as the information carrying quantity, the signal-to-noise ratio that is maximally attainable for a resistor equals:

$$\frac{S}{N} = \frac{v_s^2}{4kTBv_s/I} = \frac{v_s I}{4kTB}, \tag{3.10}$$

where v_s is the given signal level and B is the bandwidth. The signal-to-noise ratio is lower for a lower current consumption.

If the current is chosen to be the signal carrying quantity, the signal-to-noise ratio is given by:

$$\frac{S}{N} = \frac{I^2}{4kTB/R} = \frac{v_s I}{4kTB}, \tag{3.11}$$

when it is assumed that the voltage swing is not changed. This expression is the same as the equation for the voltage-mode case: the signal-to-noise ratio also reduces for reducing current consumption. However, for this situation an additional degree of freedom is available: the voltage swing. When the voltage swing is increased by the factor by which the current consumption is reduced, the signal-to-noise ratio does not change.

The bipolar transistor For the bipolar transistor three noise sources are predominantly responsible for its noise performance:

- collector shot noise;
- base shot noise;
- thermal noise of the base resistance.

These three sources contribute to the noise performance of a circuit. The relative contribution is determined by the collector bias current and the source impedance [2]. In the first-order approximation it does not matter whether it concerns a voltage source or a current source.

Consider the input configurations as depicted in figure 3.11. Both input configurations are a part of a complete amplifier; the two controlled sources, i_{fb} and v_{fb} represent the current and voltage feedback respectively. When the feedbacks are assumed to be ideal, the noise contribution of the bipolar transistor is equal for both situations. This is easily verified with the Norton-Thévenin equivalents. For both configurations the noise level is minimal when it holds that:

$$I_{C_{opt}} = \frac{V_T\sqrt{\beta}}{R_s + r_b}, \tag{3.12}$$

where V_T is the thermal voltage and r_b is the base resistance. The signal-to-noise ratio for both configurations equals for $I_{C_{opt}}$:

$$\frac{S}{N} = \frac{v_s^2}{4kT(\frac{R_s+r_b}{\sqrt{\beta}} + r_b + R_s)B}, \tag{3.13}$$

[2] The source impedance should be interpreted in a wide sense. It is the impedance from which the transistor is driven, i.e. signal-source impedance, driving-stage impedance, et cetera.

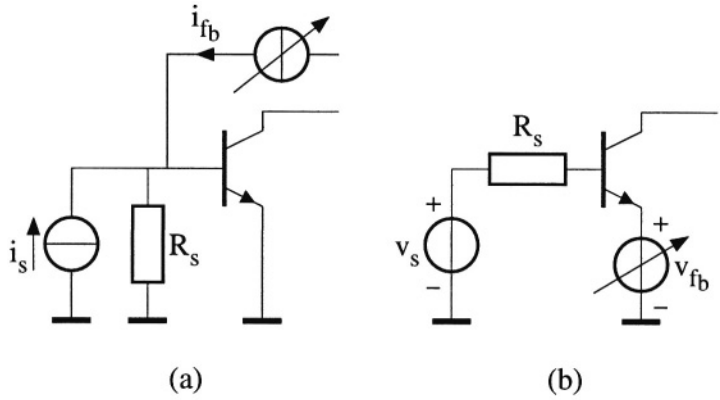

Figure 3.11: Input configurations: a) current input and b) voltage input.

where B is the bandwidth and v_s and i_s are related according to $v_s = i_s R_s$.

Thus for the *noise* it makes no difference whether the voltage or the current is used as the signal source. The optimum collector current depends on the source *impedance* and not on the source *type*. When reducing the collector current in order to reduce current consumption, the noise level always increases. The original current level was found from a noise minimization, and thus lowering the current means a non-optimal collector current.

The field-effect transistor The equivalent input noise sources of the field effect transistors (MOSFET and JFET) are substantially different with respect to the equivalent input noise sources of the bipolar transistor. For the bipolar transistor the equivalent input noise voltage increases for reducing collector current, whereas the equivalent input noise current reduces for reducing collector current. Therefore an optimum value can be found, not being a boundary value, for which the noise level is minimal for a given source impedance. For FET-like components the DC gate current can be assumed to be zero and, consequently, no shot noise is introduced by the gate junction. When the noise of the drain of the FET-like components is transformed to the input, an equivalent voltage and current noise source are found. For both sources it holds that their value increases for decreasing drain current.

Again, the SNR is independent of the type of input signal in the first-order approximation. For the signal-to-noise ratio at the input of an amplifier with a FET as input component, it holds that:

$$\frac{S}{N} = \frac{v_s^2}{\int_0^B \{4kTc/g_m\,[1 + \omega^2 R_s^2 (C_{GS} + C_{GD})^2]\}\,d\omega}, \tag{3.14}$$

where g_m is the transconductance of the FET, C_{GS} and C_{GD} are the gate-source

and gate-drain capacitance, respectively, and R_s is the source impedance. The SNR increases for increasing g_m which can be done by enlarging the drain current. The maximum is obtained for the maximum permissible current for which g_m is a maximum. But, as the g_m of a FET (proportional to the square root of the drain current in the strong-inversion region and proportional to the drain current in weak-inversion region) is lower for a given current than the g_m of a bipolar transistor [16] (proportional to the collector current), the *SNR* with a bipolar input is very often higher. Only in the case of high input impedances may the FET be favorable. The optimum value of the collector current becomes very low, see equation (3.12) and consequently the behavior of the transistor may degrade considerably. In that case the FET may cope with this problem as its optimum is found at the maximum drain current.

In the case of a MOSFET biased in the sub-threshold region, the output noise can be considered to be analogous to the noise of the bipolar transistor, i.e. $2qI_D$ [16] and the FET almost resembles the noise behavior of the bipolar transistor and can even become better in other aspects.

3.4.3.2 Signal power

For the maximum output signal power it does not matter whether a voltage or a current is used. For the maximum signal supplied to a load, the following equation holds:

$$V_{max} = \hat{I}|Z_{load}|, \tag{3.15}$$
$$I_{max} = \hat{I}, \tag{3.16}$$

where $|Z_{load}|$ is the modulus of the load impedance and $\hat{I}$ is the maximum allowed current consumption. Both maximum signals are proportional to $\hat{I}$. In both cases the maximum power supplied to the load equals:

$$P_{max} = \hat{I}^2|Z_{load}|. \tag{3.17}$$

Especially in the case of the output of an amplifier, where the signal currents are mostly the largest, it is important that the output stage is current efficient. That means that the ratio of the bias current and the signal current is as small as possible. Ultimately this means that the output stage is class-B biased. To prevent there being an unacceptable distortion level, due to cross-over distortion, for instance, class-AB biasing may be used.

3.4.3.3 Bandwidth

As was found from the preceding discussions, the impedance level of a circuit increases as a result of the reduced current consumption. Consequently, parasitic impedances may exert more influence. In figure 3.12 two different types

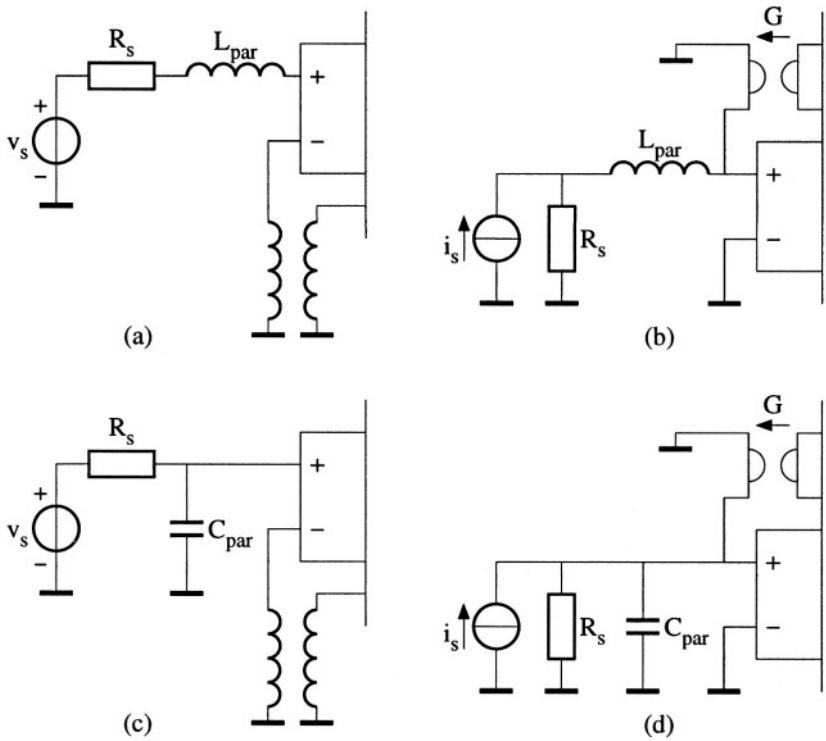

Figure 3.12: A voltage and a current input each with a parasitic capacitance or a parasitic inductance.

of feedback are depicted, each for two different types of source impedance. For the feedback networks the ideal network-theoretical elements are used. For configuration a, the current flowing through the voltage source is ideally zero and, consequently, the series inductance does not influence the accuracy of the transfer. For configuration d, the dual holds; the voltage across the source impedance is zero and thus the parasitic capacitance does not degrade the accuracy of the transfer.

In contrast, for configurations b and c the parasitics may exert an influence on the accuracy of the transfer. For configuration b the inductance causes a pole in the ideal transfer at:

$$p = \frac{-R_s}{2\pi L}, \tag{3.18}$$

for which the active part is assumed to be a nullor. Especially for the relatively low output impedances of the current source, this pole may become noticeable. Configuration c has a pole in its ideal transfer at:

$$p = \frac{-1}{2\pi R_s C}, \tag{3.19}$$

which is especially noticeable for relatively high output impedances of the voltage source.

For low-current integrated circuits the predominant parasitics are the capacitors and the impedance levels are relatively high. Considering the input of a signal-processing path, current processing may be favorable as it can be done more *accurately,* the predominant parasitic impedances (the capacitors) are incorporated in a negative-feedback loop.

For the output of a signal-processing path an analogous discussion holds. From that, it follows that a voltage at the output may be favorable as the parasitic capacitance is then enclosed by the negative-feedback loop, whereas it results in a pole, which is directly in the transfer, for the current processing. Thus the optimum overall transfer type is the transimpedance type.

For the intermediate interfaces it does not matter whether the current or voltage is used. In the case of voltages, the output of the driving block incorporates the parasitic capacitance in its loop, and for the currents the driven block encloses the parasitics in its loop.

So, it may appear that the choice of whether to use current or voltage does not make any difference. However, there is a difference inside the nullor implementations. As the parasitic capacitances are almost always to the substrate, which is a signal ground, it is preferable to have at least one terminal of the nullor grounded, as otherwise a fifth terminal is introduced which can hamper the frequency behavior considerably. For instance, consider a transconductance amplifier in which both ports are floating, see figure 3.13. The input impedances of the separate stages $(r_\pi s)$, are relatively low. The parasitic capacitance C_{par}

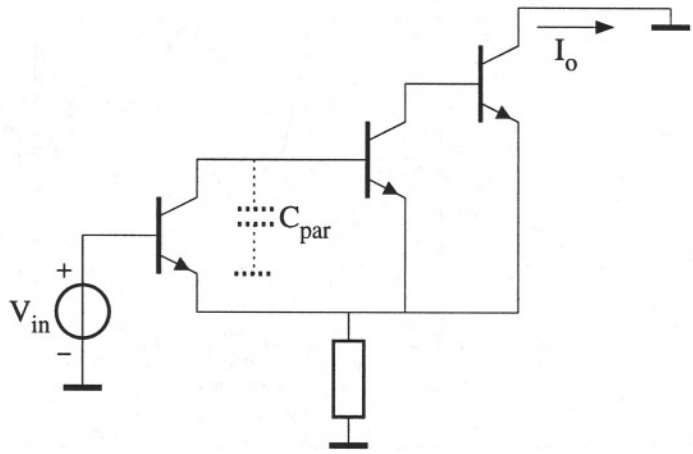

Figure 3.13: A three-stage transconductance amplifier.

however, is not completely parallel to the input of a stage, but only connected to one terminal of a stage. The other side is connected to the substrate. This causes the parallel impedance to which the parasitic is connected to be relatively high due to the series-feedback action of the total amplifier. Consequently, the influence of the parasitic is considerably enlarged.

As the transconductance amplifier is the only one having two floating ports, it should preferably not be used when maximum bandwidth is important. The other three configurations do not need to have this problem.

For the implementation of the nullor, in [6] and [8] is shown that the CE (CS) stage is the best choice. For this type of stage, in the case of grounded emitter (source), the parasitics are in parallel to the input or the output of the stage and consequently have the lowest influence on the bandwidth of an amplifier. When local series feedback is used, parasitics may exert more influence due to the higher impedance levels (cf. the example of the three-stage transconductance amplifier).

Besides the parasitics which exert a larger influence on the bandwidth of an amplifier, the maximum speed of the constituent devices itself also reduces and causes the amplifier to have a reduced bandwidth. The devices discussed here are the bipolar transistor, the field effect transistor and the resistor.

Bipolar transistor In equation (3.8) the effective transit frequency of a bipolar transistor was given. It is repeated here:

$$f_{T_{eff}} = \frac{g_m}{2\pi(g_m T_F + C_{je} + C_{jc} + C_{js})}. \tag{3.20}$$

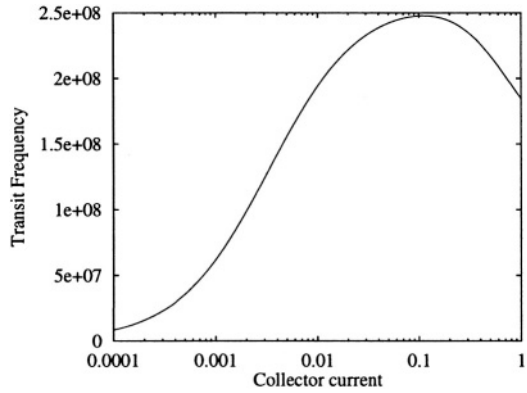

Figure 3.14: The $f_{T_{eff}}$ of a transistor as functions of its collector current.

The term $g_m T_F$ is the diffusion capacitance of the bipolar transistor, it accounts for the speed limit that is caused by the time charge requires for crossing the base of a transistor. The larger the current is, the more charge is present in the base and the larger the equivalent capacitance is. Thus for relatively high collector currents, the $f_{T_{eff}}$ is mainly determined by the diffusion capacitance and is given by:

$$f_T = \frac{1}{2\pi T_F}, \tag{3.21}$$

which is independent of the collector current and equals the transit frequency of a process. For even higher currents, $f_{T_{eff}}$ reduces again [13]; however in the case of low-current design this point is not reached. For relatively low currents, the charge in the base can be ignored with respect to the charges found in the junctions capacitances and $f_{T_{eff}}$ is given by:

$$f_{T_{eff}} = \frac{g_m}{2\pi(C_{je} + C_{jc} + C_{js})}, \tag{3.22}$$

and reduces for lower collector currents. This behavior is drawn in figure 3.14. For relatively low currents, the $f_{T_{eff}}$ of a transistor is determined by the junction capacitances. These junction capacitances can be reduced by reducing the size of the complete transistor. Capacitances reduce as the effective areas of the three corresponding junctions reduce.

When the isolation of the transistors is done by deep high-doped isolation walls, the substrate capacitance can be considerable; values can be on the order of some tenth of pF, whereas the base-emitter and base-collector capacitances are about 0.05 pF to 0.1 pF; then the substrate capacitance predominantly determines the $f_{T_{eff}}$ of the transistor. When oxide walls for the isolation are used,

the total substrate capacitance can be considerably reduced as the capacitance of the oxide wall is almost nullified. The substrate capacitance is now determined by the capacitance of the buried layer to the substrate. As the isolation wall is now oxide it is allowed to touch the buried layer and consequently the transistor size can be considerably reduced and the substrate capacitance reduces even further. When the high-doped isolation wall touches the high-doped buried layer, a junction is formed which is highly doped at both sides. Consequently, the breakdown voltage can be as low as just a few volt. For low-voltage circuits (1 V) this does not need to be a problem and this can thus be used to reduce the parasitic substrate capacitance when no oxide isolation is available.

The (MOS)FET The effective transit frequency, $f_{T_{eff}}$, of the (MOS)FET can be defined as [16]:

$$f_{T_{eff}} = \frac{g_m}{2\pi(C_{GS} + C_{GB})}, \tag{3.23}$$

where C_{GS} is the gate-source capacitance, C_{GB} is the gate-bulk capacitance and g_m is the transconductance of the (MOS)FET. The gate-source capacitance depends on the mode of operation of the MOSFET. For moderate and strong inversion it is about 60 % of the gate-oxide capacitance whereas it reduces to almost zero for weak inversion.

The (MOS)FETs are self-isolating devices and, consequently, they do not have the corresponding relatively large substrate capacitance as the bipolar transistor does. The parasitic capacitances to the bulk are found at the drain and source contact. A part of these parasitics is caused by the channel-to-bulk capacitance for the MOSFET.

For the JFET it depends on the type whether the parasitic from the channel to the bulk is absent or not. When the gate of JFET completely encloses the channel, the channel is isolated from the bulk and consequently the parasitic capacitance from the channel to bulk is zero. The parasitic from the gate to the bulk can be relatively large now, as the gate-bulk junction now has a relatively large area.

The JFET is, compared with the MOSFET and bipolar transistor, a rather bulky element and, consequently, relatively slow. Its application is therefore mostly in the field of low-frequency controlled resistance.

Resistors The bandwidth of resistors also reduces when a low current consumption is required. When currents are in the sub-μA region and the voltages required across resistors are still in range of several volt, resistances become very high. When these resistances are realized on chip, the resistors get relatively large and, consequently, the parasitic capacitances to the substrate are considerable. Analogous to interconnects on a chip, these relatively large resistors

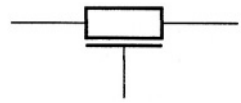

Figure 3.15: The resistor visualized as a distributed RC network.

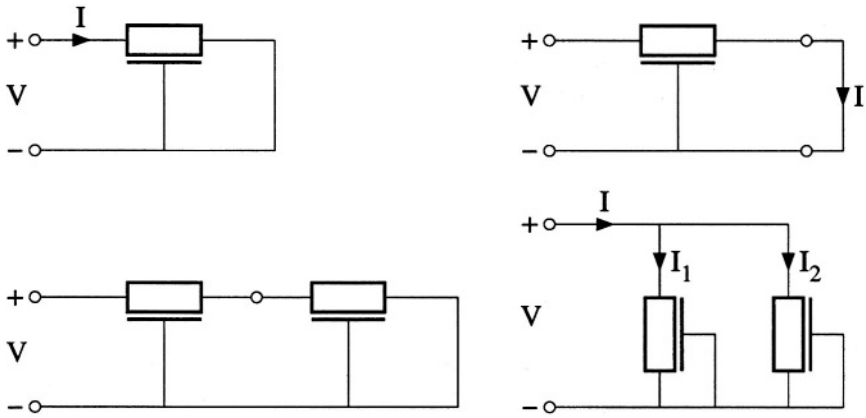

Figure 3.16: The four ways in which resistors can be used.

can be modeled by a distributed RC network, see figure 3.15. In the literature a lot has been published concerning approximations of distributed networks in order to predict the step-response of an interconnect modeled as a distributed RC network [17], [18] and [19]. The approximations are based on time-domain considerations. Analog design, however, is preferably done in the frequency domain and does not use the step-response but instead uses phase and amplitude of signals. Therefore, these approximations are not the appropriate ones. Other publications deal with frequency domain approximations [20]; however, these approximations assume a frequency much lower than the first pole. For low-current applications this does not need to be true. Therefore, some attention is paid here to the frequency behavior of high-ohmic resistors modeled as distributed RC networks. The analysis is based on the calculations done by Deily [18] and Zurada [21].

Resistors can be used in four ways, see figure 3.16:

- one-port configuration V/I;
- two-port configuration I/V;
- two-port configuration, voltage divider;
- two-port configuration, current divider.

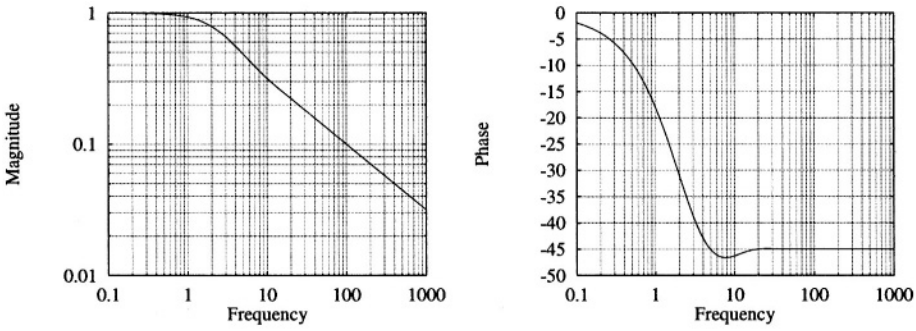

Figure 3.17: The Bode plots of a resistor, used as a one port, with a distributed parasitic capacitance.

The current divider can be seen as the resistor as one-port for which the port impedance is important and is therefore not separately analyzed. The influence of the distributed parasitics is different in the other three situations. For the one-port configuration, the impedance can readily be found as:

$$Z_R(s) = R \cdot \frac{\sinh\left(\sqrt{sRC}\right)}{\sqrt{sRC}\cosh\left(\sqrt{sRC}\right)}, \tag{3.24}$$

where s is the Laplace variable, R is the DC resistance and C is the total parasitic capacitance. Calculating the poles and zeros, one ends up with the following expression:

$$Z_R(s) = R \cdot \frac{\prod_{n=1}^{\infty}[s\frac{RC}{n^2\pi^2}+1]}{\prod_{n=0}^{\infty}[s\frac{RC}{(n+\frac{1}{2})^2}+1]}. \tag{3.25}$$

This impedance has an alternating sequence of poles and zeros on the negative real axis. The bode plots of this impedance are depicted in figure 3.17. From the equations follows that the pole with the lowest frequency can be found at:

$$p_{first} = \frac{-\frac{1}{4}\pi^2}{2\pi RC}, \tag{3.26}$$

in which the magnitude is about 2.5 times higher than when the time constant RC is used. The corresponding -3 dB point also follows from the plot [3]. A closer look at the plot shows that the roll-off equals 10 dB/dec and the phase shift limits to 45°. This is result of the alternating sequence of poles and zeros.

[3] It is permissible to say that the pole of equation (3.26) determines the -3 dB point as the magnitude of the next pole is a factor 9 higher and of the first zero is a factor 4 higher.

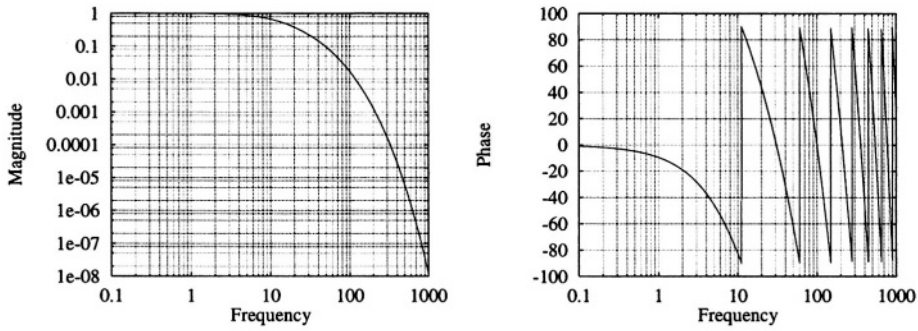

Figure 3.18: The bode plots of a resistor used as a two port.

When the resistor is used as a two-port V-to-I converter, the following transfer can be found:

$$\frac{I_{out}}{V_{in}} = G \cdot \frac{\sqrt{sRC}}{\sinh\left(\sqrt{sRC}\right)}. \tag{3.27}$$

Again, calculating poles and zeros the following expression is found:

$$\frac{I_{out}}{V_{in}} = G \cdot \frac{1}{\prod_{n=1}^{\infty}(s\frac{RC}{n^2\pi^2}+1)}. \tag{3.28}$$

The impedance has a sequence of infinite countable poles on the negative real axis and the one with the lowest frequency can be found at:

$$p_{first} = \frac{-\pi^2}{2\pi RC}. \tag{3.29}$$

The bode plots are depicted in figure 3.18. For this situation the order of the conductance steadily increases as a function of the frequency, resulting in a phase shift not limited to 90° but steadily increasing for increasing frequency.

The third situation is in which the resistors are used as voltage divider. This voltage division cannot be derived from the impedance found for the one-port resistor by calculating a voltage division with $Z_1/(Z_1 + Z_2)$, because for both the resistors the parasitic capacitances are connected to the substrate. As one of the resistors has no terminal connected to the small-signal ground (which is also the substrate), the influence of the parasitics is different for both resistors. For calculating the transfer to the output of the voltage divider, the complete divider has to be visualized as a single distributed network with a tap as the output of the voltage divider. The transfer of this network can be written as:

$$\frac{V_{out}}{V_{in}} = \frac{\sinh\left(\delta\sqrt{sRC}\right)}{\sinh\left(\sqrt{sRC}\right)}, \tag{3.30}$$

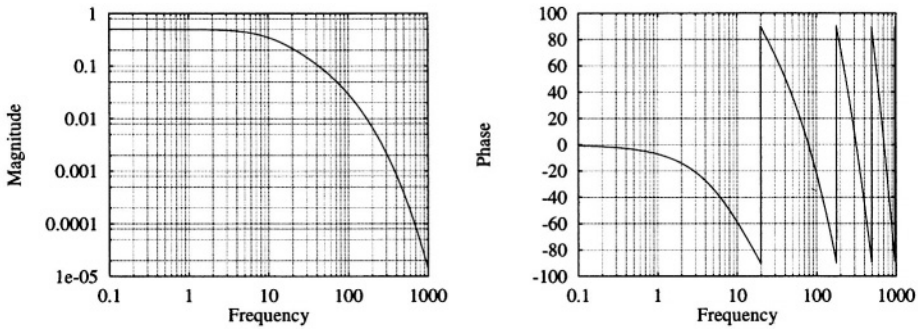

Figure 3.19: The transfer of a distributed-network voltage divider.

where δ is the voltage ratio. For the poles and zeros, the following expression is found:

$$\frac{V_{out}}{V_{in}} = \delta \frac{\prod_{n=1}^{\infty}(s\frac{\delta^2 RC}{n^2\pi^2}+1)}{\prod_{n=1}^{\infty}(s\frac{RC}{n^2\pi^2}+1)}. \tag{3.31}$$

For $\delta = 1$ the poles and zeros cancel, which may be expected as the input and output are then completely in parallel. The Bode plots of this transfer are depicted in figure 3.19. Again, the roll-off increases for higher frequencies, however, less fast than for the previous case. The same holds for the phase, it is steadily increasing for higher frequencies. The difference with respect to the resistor as one port is that the zeros and poles do not alternate. The axis on which the zeros are located, compared with the axis on which the poles are, can be said to be scaled by $1/\delta^2$. Therefore, the phase is not limited as for the one-port situation, however it does not drop as fast as for the two-port V-to-I transfer, as the zeros have completely vanished for that transfer.

In the previous expressions the capacitances due to a contact were ignored as the size of the resistors was assumed to be relatively large. The influence of the contact parasitics can be studied by adding them to the resistor and using expressions for capacitive loaded distributed networks; see, for instance, [22]. It was also assumed that one of the terminals of the networks was grounded. If the resistive networks are floating, the influence of the parasitics will increase as was seen in section 3.4.3.3.

From previous consideration it follows that when the resistor is used as a two port, the phase and frequency behavior can be very inconvenient. When these high resistances are required, the resistor may be replaced by a MOSFET or JFET of which the channel is almost depleted. The size of these components is considerably smaller and the resulting influence of the parasitics is at relatively high frequencies [20]. Also, active structures may be used at the cost of an increased current consumption; for example, the method as depicted in figure

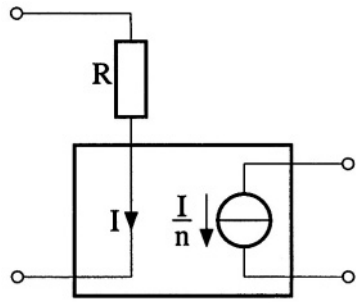

Figure 3.20: An active realization of a high resistance.

3.20 [23]. The voltage is converted into a current by a relatively low resistance *R*. The resulting current is attenuated by means of a current attenuator and the resulting impedance equals:

$$R_{eq} = R * n, \tag{3.32}$$

where n is the attenuation factor.

3.5 Low power

3.5.1 Minimization

Low-power design means *both* low-current design *and* low-voltage design. Therefore, low-power design has to do with the combined difficulties of low-voltage and low-current design. On top of that, when assuming low-power design, the orthogonality is terribly disturbed. As was seen all the quality aspects of electronic circuits improve when more current is allowed to be consumed. Now one has to find answers on questions as: do I use an additional μW to improve the noise performance? or: do I use this power to improve the bandwidth of my circuit? To be able to answer this question, at least a weighting function must be available which relates noise performance and bandwidth to parameters having the same dimension such that they can be compared objectively. In this weighting function factors are found representing subjective aspects such as, for instance, the intelligibility of an audio signal is larger when the bandwidth is enlarged than when the noise level is reduced. Besides these subjective weighting factors, which depend on the application and consumer, also objective weighting factors are found, for instance, when the current of the input stage is increased by 1 μA, the noise level only reduces by 0.1 dB whereas the bandwidth increases by 50%. This type of factors is independent of the situation and can be coped with in general design methods.

Consequently, the strategy should be that first the separate quality aspects are optimized assuming orthogonality and in the second step power reduction and/or power exchange between the several parts of the circuit can be done such that the performance of the circuits reaches the requirements.

A straightforward way to reduce the power consumption is by reducing the power-supply voltage and/or by reducing the current levels. Less straightforward is to exchange voltage and current, such that the power *efficiency* is enlarged. This is discussed in the next section.

3.5.2 Power-supply voltage conversion

The starting point is the assumption that the principal power source has a relatively low voltage, i.e. 1 V. It may then be that a circuit block requires a higher supply voltage in order to function in an optimum way. These minimally required node voltages can be found via systematic biasing techniques as presented in [8].

Realizing higher voltages from a fixed lower supply voltage can be done by means of a DC-DC converter. Two different types can be distinguished:

- Charge-pump based DC-DC converters;
- LC-tank based DC-DC converters.

Charge-pump based DC-DC converters Of this type the three main configurations are [24]: the Marx voltage multiplier [25], [26], the Cockcroft-Walton voltage multiplier [27] and the Dickson voltage multiplier [28]. These voltage multipliers are based on charge pumping by means of capacitors and self-timed switches. Depending on the implementation of the self-timed switches, an efficiency on the order of 80% can be obtained with MOS switches and an efficiency of about 40 % is possible with bipolar switches. The main difference is that in the on-state the voltage drop across the MOS switch can be considerably lower than for the bipolar switch [24].

LC-tank based DC-DC converters The more commonly used name for this type is the Buck converter. These use an LC tank to convert the voltage, [29]. For this type of converters only one switch at the input side is found and, therefore, does not need to be self-timed. Using bipolar switches may also result in a high efficiency, i.e. on the order of 90 % and higher. The disadvantage of this type of converters is the inductor required, very often making an external component inevitable.

Concluding, the use of voltage multipliers to increase the supply voltage (locally) is probably only effective in improving the overall performance when the output power of the voltage multiplier is relatively low (power efficiency).

An application could be the generation of a higher voltage in order to drive the gate of (MOS)FETs to a voltage which is well enough beyond the drain voltage (which is at most the supply voltage) in order to gain some extra drain voltage swing, [30] .

3.6 Conclusion

This chapter described the influence on the performance of electronics when low-voltage low-power constraints are given. To get a clear insight, these constraints were split into three separate parts:

- low voltage;
- low current;
- low power.

Low voltage was shown to mainly influence the maximum possible signal voltages. For a lower supply voltage, the maximum value of the voltage signals reduce. When using the current as the information carrier, an additional degree of freedom is obtained, i.e. the impedance level, to set the maximum current level. Further, it was shown that the nullor implementations are hardly influenced. The minimum required voltage room is only slightly larger than the minimum required voltage room for the constituent devices. For a floating port, the minimum required voltage is a saturation voltage of a current source larger in comparison to the non-floating port, thus a slight preference may be found for the nullor implementation with non-floating ports.

The performance of current sources was shown to reduce for lowering the supply voltages whereas the performance of voltage sources improves or, at least, remains the same.

The devices are hardly influenced either, only the junction capacitances may be slightly larger, resulting in a speed reduction. But, it may be said to be negligible.

Low current was shown to have a predominant influence on the signal power, bandwidth and noise performance of devices. A reduction in the current consumption means a reduction in the performance of all three design aspects of the resistor, bipolar transistor and the (MOS)FET. It does not matter whether the voltage or current is used as the information carrying quantity.

Low power is the combination of low voltage and low current. To make optimal use of the available power, voltage and current have to be interchangeable.

Each part of the circuit should be supplied from a voltage source with the minimum required value. This is possible with voltage multipliers. However, due to the low input voltage, i.e. 1 V, the efficiency of these multipliers is moderate to low, or an external inductor is required to obtain better efficiency. Therefore, for low-voltage low-power design the use of voltage multipliers is limited to those parts of the circuits with a relatively low-power consumption, i.e. gates of (MOS)FETs.

Due to the low-power constraint, the orthogonalization of a design procedure is hampered. Now first the separate blocks have to be optimized, independently, and second a weighting must be used for the allowed power consumption of the separate blocks.

Thus an overall conclusion is that due to the low-voltage constraint, signal voltages (related via impedance levels to the current signals or directly being the signals) are limited to the supply voltage, which is very trivial. Low current reduces the performance with respect to noise, signal power and bandwidth. Low power combined with low voltage may hamper the orthogonality in the design process but does not introduce any additional performance degradation.

Bibliography

[1] R.J. Widlar. Low voltage techniques. *IEEE Journal of Solid-State Circuits,* 13(6):838–846, December 1978.

[2] H.R. Camenzind and R.B. Kash. A low-voltage IC timer. *IEEE Journal of Solid-State Circuits,* 13(6):847–852, December 1978.

[3] J. Fonderie. *Low-Voltage Bipolar Operational Amplifiers.* PhD thesis, Delft University of Technology, November 1991.

[4] W.A. Serdijn. *The Design of Low-Voltage Low-Power Analog Integrated Circuits and Their Applications in Hearing Instruments.* PhD thesis, Delft University of Technology, February 1994.

[5] C. Toumazou, F.J. Lidgey, and D.G. Haigh, editors. *Analogue IC Design: The Current-Mode Approach.* Peter Peregrinus, London, 1990.

[6] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[7] E.A. Vittoz. Low-power design: Ways to approach the limits. In *Proceedings of the IEEE International Solid-State Circuits Conference,* pages 14–18, February 1994.

[8] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft

University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

[9] J. Bardeen and W.H. Brattain. The transistor, a semi-conductor triode. *Physical Review,* 74:230–231, June 1948.

[10] A.C. van der Woerd and A.C. Pluygers. Biasing a differential pair in low-voltage analog circuits: A systematic approach. *Analog Integrated Circuits and Signal Processing,* 3:119–125, 1993.

[11] A.C. Pluygers. A novel microphone preamplifier for use in hearing aids. *Analog Integrated Circuits and Signal Processing,* 3:113–118, 1993.

[12] A. van Staveren and A.H.M. van Roermund. Low-voltage low-power controlled attenuator for hearing aids. *Electronics Letters,* 29(15):1355–1356, 1993.

[13] I.E. Getreu. *Modeling the Bipolar Transistor.* Elsevier, New York, 1978.

[14] P.J.M. van Adrichem. *Design Manual DIMES-01 Process.* Technical University of Delft and Delft Institute for Micron and Sub-micron Technology (DIMES), December 1993.

[15] A. van Staveren, G.L.E. Monna, C.J.M. Verhoeven, and A.H.M. van Roermund. A low-power class-ab negative feedback amplifier for a 1V LW receiver. *Analog Integrated Circuits and Signal Processing,* 20:63–75, 1999.

[16] Y.P. Tsividis. *Operation and Modeling of the MOS Transistor.* McGraw-Hill, 1987.

[17] J. Kelly and M.S. Ghausi. On the effective dominant pole of the distributed RC networks. *Journal of the Franklin Institute,* 279(6):417–429, June 1965.

[18] G.R. Deily. Closed-form solutions for voltage-step response of open and shorted distributed RC lines. *IEEE Transactions on Circuits and Systems,* 22(6):534–541, June 1975.

[19] R.J. Antinone and G.W. Brown. The modeling of resistive interconnects for integrated circuits. *IEEE Journal of Solid-State Circuits,* 18(2):200–203, April 1983.

[20] B.X. Shi, J. Khoury, and Y.P. Tsividis. High frequency effects in MOSFET-C Tow-Thomas biquads. *IEEE Transactions on Circuits and Systems,* 33(7):648–651, July 1986.

[21] J.M. Zurada and T. Liu. Equivalent dominant pole approximation of capacitively load VLSI interconnection. *IEEE Transactions on Circuits and Systems,* 34(2):205–207, February 1987.

[22] M.T. Abuelma'atti. Multi-pole approximation of capacitively loaded VLSI interconnection. In *IEE Proceedings, Part G,* volume 136, pages 118–120, June 1989.

[23] G.L.E. Monna. *Design of Low-Voltage Integrated Filter-Mixer Systems.* PhD thesis, Delft University of Technology, September 1996.

[24] M. Berkhout. *Audio Amplifiers in BCD Technology.* PhD thesis, University of Twente, October 1996.

[25] E. Marx. Investigations in the testing of insulators with impact voltages. *Electrotechniser Zeitung,* 45:652, 1924.

[26] E.A. Richley. Marx generator for high voltage experiments. *Electronics & Wireless World,* 93:519–523, May 1987.

[27] J.D. Cockcroft and E.T.S. Walton. Experiments with high velocity positive ions. Further developments in the method of obtaining high velocity positive ions. In *Proceedings of the Royal Society of London A,* volume 136, pages 619–630, 1932.

[28] J.F. Dickson. On-chip high-voltage generation in NMOS integrated circuits using an improved voltage multiplier technique. *IEEE Journal of Solid-State Circuits,* 11(3):374–378, June 1976.

[29] M. Brown. *Practical Switching Power Supply Design.* Motorola, Series in Solid-State Electronics. Academic Press Inc., San Diego, 1990.

[30] G.L.E. Monna et al. Charge pump for optimal dynamic range filters. In *Proceedings of the IEEE International Symposium on Circuits and Systems,* volume 5, pages 747–750, May 1994.

Chapter 4

Amplifiers

4.1 Introduction

Amplifiers are required in almost every electronic system. In measurement equipment at the beginning of an information processing chain, they are required to make the, very often, relatively weak information signals more robust so that they are more easily processed and transmitted. At the end of the information processing chain the power level of the information signal requires power amplification in order to be able to drive the output transducer. Each of these amplifiers is not allowed to deteriorate the information signal.

In the world of amplifier design two trends can be distinguished, viz. the design of:

- general purpose amplifiers (the opamps);
- dedicated amplifiers.

The main difference between those two types of amplifiers is that the general purpose amplifier has to be applicable for a wide range of load conditions and feedback factors, whereas for dedicated amplifiers the load is relatively well known. Therefore, the opamps are frequency compensated such that, approximately, a first-order frequency behavior is obtained [1], [2], so that the risk of instability due to load variations is reduced to a large extent. In contrast, for dedicated amplifiers the order of the frequency behavior can be larger than one as the load is better known. However, orders beyond three become practically almost impossible as these amplifiers can become instable due to a relatively small additional phase shift.

In this chapter, the amplifier is assumed to be a dedicated amplifier as it is to be used in the oscillator or bandgap reference by which the load and feedback conditions are relatively well known.

4.2 The basic function

The basic function of the amplifier is to accurately change the power or the signal level and/or the dimension of an information signal, e_{in}:

$$\frac{e_{out}}{e_{in}} = H_{transfer}, \tag{4.1}$$

where e_{out} is the amplified signal and $H_{transfer}$ is the transfer of the amplifier. To be able to reach the required accuracy, devices with an accurate transfer have to determine the overall transfer of an amplifier. In chapter 3, the asymptotic-gain model [3] was shown to be the appropriate model to synthesize accurate amplifiers. This model assumes a linear accurate feedback network (resistors, for instance) and an active part, supplying the required power gain which does not need to be accurately specified. When the loop gain of the amplifier is large, the transfer of the amplifier is determined by the feedback network and an accurate amplification is obtained.

This model can only be used for linear amplifiers as it is based on the superposition principle. The model assumes that by means of a relatively large loop gain, the non-linearity of the active part is counteracted and a linear approximation can be used, i.e. the small-signal equivalent.

Recently, intrinsic non-linear electronics is gaining more interest and is believed to be a serious candidate to use the high frequency potentials of, for instance, SiGe processes. Due to the intended non-linearity, signals are spread over a relatively large frequency range on the chip making it less sensitive to contamination (cf. FM modulation). One could think of using the very accurate exponential relation between the base-emitter voltage and collector current of bipolar transistors, or the gate-source voltage and drain current of weak-inversion MOS transistors. These accurate non-linearities are suppressed in linear electronics by using a lot of loop gain and cause distortion when they pop up in the transfer. For the non-linear electronics these non-linearities do not need to be suppressed any longer, but instead are used favorably. Examples of this non-linear signal processing is found in exponential state space filters [4].

In principle, the amplification factor is frequency independent, it is the implementation of the mathematical scaling:

$$y = a * x, \tag{4.2}$$

where a is the scaling factor. This scaling is speed independent. Introducing speed limitations are an additional step; adding filtering, for instance.

4.3 Relation to the fundamental design aspects

An amplifier can be described as an information transporting channel. The information supplied at the input of the amplifier has to be completely found

again at the output of the amplifier. Practical amplifiers always introduce errors. The type of errors are described in chapter 2. Applying these errors to the design of amplifiers one can say that the stochastic noise sources limit the dynamic range of an amplifier at the lower side, thermal noise of the input devices, for instance. The systematic errors result in a limitation at the upper side of the dynamic range which is thus closely related to the maximal signal power that can be handled by the amplifier; too large signals result in systematic errors.

The bandwidth of an information transporting channel is directly related to the bandwidth of the amplifier. The smaller the bandwidth of an amplifier is, the less information can be transported.

These three fundamental design aspects are discussed in the following sections.

4.4 Noise

The input of an amplifier is mostly the place where signals are the smallest and thus where noise can have the largest influence. Methods for the minimization of the noise of an amplifier can be found in numerous textbooks [3], [5], [6] and [7]. The starting point for the noise minimization is the transformation of the relevant noise sources into an equivalent noise source at the input of the amplifier such that it is easily compared with the signal source.

To reduce the number of amplifying stages contributing to the equivalent input noise source and thus to keep the noise level as low as possible, input stages have to be used whose four chain parameters are relatively small [3] or in other words have high gain parameters [2]. The noise generated in the subsequent amplifying stages can then be ignored [8] and the signal-to-noise ratio is governed by the feedback network, the input stage and the source.

The dimensioning of the feedback network is given by the required transfer. However, for the voltage and current amplifier the impedance level of the feedback network can be freely chosen and can, at the cost of power consumption, be chosen such that the noise contribution is reduced to an acceptable level.

The noise contribution of the input stage is determined by its bias current and for (MOS)FETs also by the transistor size. For bipolar transistors the equivalent voltage noise and equivalent current noise of the bipolar transistor decrease and increase, respectively, for increasing collector current, therefore, a minimum can be found. For the MOS transistor, both its equivalent voltage and current noise reduce for increasing drain current. This is a consequence of the absence of the DC input current.

The relative contribution of the voltage and current noise is determined by the level of the source impedance. For a source with a low impedance, the voltage noise is accentuated in comparison to the current noise. Therefore, the optimum bias current of the input transistor is at relatively high currents. Further,

for a frequency dependent source impedance, a frequency dependent weighting function is found for the noise. For a capacitive source, the current noise is predominant for relatively low frequencies and the voltage noise is predominant for relatively high frequencies.

Thus the relevant part of the circuit, concerning the design with respect to noise, is mainly located at the input of an amplifier.

4.5 Distortion

Distortion of a signal occurs, generally speaking, when the output signal deviates from the ideal output signal, i.e. the signal which would be found at the output of the ideal amplifier. First of all, linear distortion is found when the intended *amplitude* of sine wave components differs from the intended amplitude *and* no frequency components are found other than those which are found in the input signal. This type of distortion occurs in the case of a limited bandwidth, for instance. But deviations in the gain factor as a result of spread in the feedback component(s) can also be responsible for linear distortion. This type of distortion is not orthogonal to the fundamental aspects, as discussed in chapter 2, as it describes bandwidth behavior, systematic errors in the feedback network, et cetera. This term is therefore not used any further in this book.

The non-linear distortion alters the frequency contents of the signals. In chapter 2 the distinction was made between:

- clipping distortion;
- weak distortion.

In amplifier design these two types of distortion have a specific influence on the behavior of the amplifier.

4.5.1 Minimization of clipping distortion

Clipping distortion is found when signals no longer fit between the supply rails, or the current-driving capability of a stage is not sufficient. Both phenomena result in an effectively broken feedback loop. As for an increasing input signal the output signal no longer increases, the signal which is fed back no longer changes and consequently the loop is broken. Characteristic of this type of distortion is the loss of information. The output signals of the amplifier are no longer unambiguous.

This type of distortion can be caused by each stage of the amplifier. At the output of the amplifier both clipping in the voltage domain and clipping in the current domain may occur as the output signals are mostly the largest. For the other stages clipping in the current domain is the one most likely to occur.

Slewing of stages is a well-known effect which causes this type of distortion; for higher frequencies the available *current* of a driving stage is completely used by the driven stage as a result of a capacitive input impedance. Consequently, the driving stage is switched off and the feedback loop is broken.

Clipping distortion is a very severe type of distortion. Originally the loop suppresses non-linearities by means of its loop gain and the loop is broken rather abruptly and all the distortion products are found at the output; including the distortion products caused by the clipping of the signal.

The amplifier has to be prevented from clipping signals. As the loop is broken at the moment of clipping, a large loop gain does not help to prevent the amplifier clipping. The only method, which is rather simple and straightforward, is to ensure there are sufficient current and voltage driving capabilities. These can be found for instance, from an AC analysis with SPICE [9]. In that case SPICE first determines the DC operating point of the circuit, subsequently the small-signal circuit is derived and this is used to perform the AC analysis. As no bias sources are found in the small-signal circuit, no clipping occurs and the maximum signals are free to become very large. From these simulation results, the minimum required driving capabilities can be found.

4.5.2 Minimizing weak-distortion

Weak distortion is found when the static transfer of the amplifiers deviates from the intended transfer and signal ambiguity is still present. For bipolar amplifiers two main types of causes can be distinguished:

- β distortion;
- g_m distortion.

β distortion is found in bipolar transistors when they are current driven and have a low-ohmic load. In this case β_f is the key parameter for the transfer of the transistor. As β_f is only slightly dependent on the current level, the resulting distortion is relatively low.

The g_m distortion, however, introduces more distortion. This type of distortion is found in voltage driven transistors which have relatively (compared with their output resistance) low-ohmic loads. In this case the transfer is mainly determined by the g_m of the transistor, which is inversely proportional to the collector current.

The two dual situations, i.e. relatively high-ohmic load and a high and low-ohmic drive, are less practical. In the case of the low-ohmic drive and a high-ohmic load, the voltage gain of the transistor is the key parameter. It is approximately independent of the collector current and thus no distortion is found. This situation is dual to the situation of the β distortion. However, for this situation the impedance levels are much higher and as a consequence

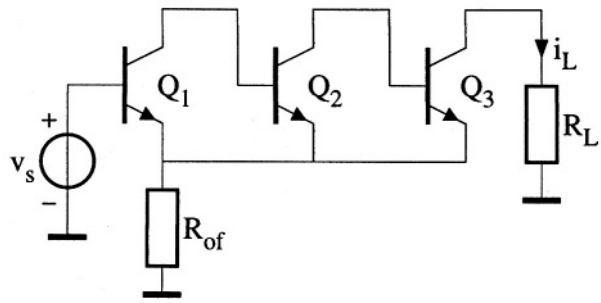

Figure 4.1: A three-stage transconductance amplifier.

the parasitic capacitances have a larger influence. Further, it is impractical to realize subsequent stages with a relatively high input impedance without the use of local feedback. For the other situation, a high-ohmic drive and a high-ohmic load, the distortion is a result of the collector-current dependence of the output resistance of the transistor. For the same reasons as the previous case, this situation is believed to be impractical.

Thus for practical amplifiers the β and g_m distortion can be assumed to be the main causes of distortion. Of course, for the highfrequencies, $f > f_T/\beta_f$, the β distortion tends to become g_m distortion as, due to the the input capacitance of the stage, the intrinsic transistor becomes voltage driven; the driving current flows mainly through the input capacitance resulting in an input voltage which results, via the g_m, in the output current.

How the g_m distortion and ultimately the β distortion can be kept to a minimum is discussed using the amplifier as given in figure 4.1. At the input a voltage comparison is realized by means of the series feedback, and thus g_m distortion is the source of distortion in this stage. Both Q_2 and Q_3 are current driven and therefore for them the β distortion is predominant. It must be noted that the predominant cause of the weak distortion is found at the *input* stage. The principal cause of the distortion is the variation of the working point of the transistors as a function of the input signal. This variation has to be reduced in order to lower the distortion. For a given input signal of a stage, the β distortion can be reduced by increasing the bias current of the stage; reducing the relative variations of the collector current. For the g_m distortion, increasing the collector current in itself does not reduce the distortion as, due to the exponential relation of the transistor, the relative variation of the collector current is given for a given input voltage. To reduce this type of distortion, the input voltage has to be reduced. This is possible when negative feedback is used. Now, by increasing the collector current, the g_m increases and as a result the loop gain increases. This leads to a smaller signal at the input of the nullor implementation. Thus, the input voltage of the stage is reduced and the

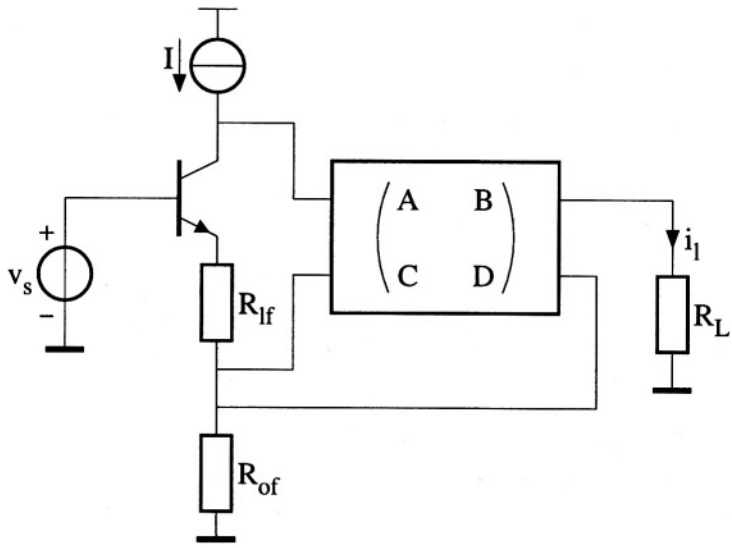

Figure 4.2: Reduction of the distortion of the input stage by means of local feedback.

relative variation of the collector current is reduced. Of course, increasing the loop gain of the amplifier also reduces the β distortion as the input currents of the corresponding stages reduce.

Thus, generally speaking, weak distortion can be reduced by increasing the loop gain. This leaves the design of the distortion, which is mainly caused in the input stage for a voltage input, orthogonal to the noise minimization.

For the MOS transistor only g_m distortion is possible. The intrinsic MOS transistor is always voltage driven. However, for MOS transistors not working in the weak-inversion mode, the g_m is approximately proportional to the square root of the drain current. Thus increasing the drain current in itself reduces the relative variation of the drain current for a given gate-source voltage. Of course, increasing the loop gain in the case of a negative-feedback amplifier is also applicable.

4.5.2.1 Using local feedback

The influence of local feedback on the distortion is studied in this section. The distortion due to the variation of the g_m may be assumed to be the worst case. To reduce this type of distortion, local feedback with R_{lf}, as depicted in figure 4.2, is introduced. The local feedback is realized at the input stage of the transconductance amplifier as this stage suffers from g_m distortion. The local

feedback is, of course, such that the voltage-to-current transfer of this stage is linearized. The remaining stages of the transconductance amplifier are described by a chain matrix with chain parameters, A, B, C and D.

The overall transfer from input voltage, v_s, to output current, i_l, equals:

$$A_t = \frac{i_l}{v_s} = \frac{1}{R_{of}} \cdot \frac{1}{1 + \frac{D}{\beta} + \frac{D}{R_{of}} \cdot \frac{V_T}{I_C} + D \cdot \frac{\beta+1}{\beta} \cdot \frac{R_{lf}}{R_{of}}}, \tag{4.3}$$

where β is the current-gain factor and I_C is the collector current of the input stage, D is the transmission parameter of the remaining stages, R_{of} is the resistor for the overall feedback, R_{lf} is the resistor for the local feedback and V_T is the thermal voltage. The transfer for the amplifier without local feedback follows from this expression by simply making R_{lf} zero. From this expression it also follows that introducing local feedback results in a reduced overall accuracy; the denominator of the second fraction of equation (4.3) deviates more from one for a $R_{lf} \neq 0\ \Omega$ compared with the situation that $R_{lf} = 0\ \Omega$. The cause is a reduced overall loop gain; the input stage is locally fed back, its transfer is reduced and, consequently, contributes less to the overall loop gain.

For finding a qualitative measure of the influence of the local feedback on the distortion, the variation of the gain as a function of the collector current can be derived. This quantity is linearly related to the differential error [10] from which second and third-order distortion can be derived very easily. The differential gain, ϵ, of a transfer A_t, is defined as:

$$\epsilon_{A_t} = \frac{A_t(e=0) - A_t(e=e_i)}{A_t(e=0)}, \tag{4.4}$$

where $A_t(e=0)$ is the quiescent transfer, i.e. without input signal, and $A_t(e = e_i)$ is the transfer in the case of an input signal e_i. For the second-order and third-order harmonic distortion, d_2 and d_3, respectively, hold:

$$d_2 = \frac{\epsilon^+ - \epsilon^-}{8}, \tag{4.5}$$

$$d_3 = \frac{\epsilon^+ + \epsilon^-}{24}, \tag{4.6}$$

where ϵ^+ and ϵ^- are the differential gain for a positive and negative signal excursion, respectively. For low distortion the gain must be as independent as possible of the signals, i.e. the differential-gain should ideally be zero.

For the variation of the transfer as a function of the input signal, it is assumed that the output current of the amplifier is the same with and without local feedback. As the collector signal current is related to this output current via parameter D of the two-port of the amplifier, the variations in the collector

current are equal for both situations. Therefore, the ratio:

$$\frac{\left(\frac{dA_t}{dI_C}\right)_{with}}{\left(\frac{dA_t}{dI_C}\right)_{without}} = \left(1 + \frac{D}{\beta} \cdot \frac{R_{lf}}{R_{of}}\right)^{-2}, \tag{4.7}$$

is a measure of the influence on the distortion. This expression is found using the following assumptions: $\beta/D \gg 1$, $\beta \gg 1$ and $R_{of} \gg D/g_m$, which are already inherently satisfied for relatively low overall loop gains. In this expression the ratio $(\beta/D)^{-1}$ is present which is equal to one over the maximal loop gain of the amplifier, i.e. without local feedback and an overall feedback impedance which is infinite. Thus it may be clear that only a very small reduction of the distortion is found, i.e. a few percent for relatively low overall loop gains and even lower for high overall loop gains.

For this derivation it was assumed that the magnitude of the overall loop gain remained the same. However, due to the local feedback the overall loop gain reduces slightly and consequently the overall transfer reduces slightly. Thus for the same input signal, the amplifier with the local feedback has the smallest output signal and thus the corresponding collector current variation is lower, resulting in less distortion.

Thus, due to local feedback at the input stage the resulting distortion reduces due to:

- linearization of the input stage;
- reduction of the overall transfer,

of which the latter is likely to have the largest result, i.e. the denominator of equation (4.3) shows a first-order dependence on the term (R_{lf}/R_{of}), whereas equation (4.7) shows a dependence of order minus two.

But, to get a fair comparison, the input signal of the amplifier has to be enlarged to end up with the original output signal. Then, a net increase is found in the distortion as a result of the aforementioned dependencies.

Further, the discussion up to now only focused on the distortion of the *input* stage; the local feedback resulted in a reduced overall loop gain. Consequently, the other sources of distortion are less suppressed and the distortion level may increase even further.

Moreover, when the influence on the noise performance is investigated, one has to conclude that local feedback does not improve the dynamic range of an amplifier. The equivalent noise voltage at the input of the amplifier is increased by the contribution of the local-feedback resistor, which has to be considerably larger than the overall-feedback resistor (see equation (4.7)), and by an increased influence of the noise of the second stage as the voltage-to-current transfer of the input stage is reduced.

Concluding, weak distortion is best reduced by increasing the overall loop gain.

4.6 Bandwidth

In the previous sections the design of the noise and the distortion behavior of the amplifier was discussed. The noise performance of an amplifier is located at its input, clipping distortion at its output and weak distortion is governed by the overall loop gain.

In this section it is shown that the bandwidth of an amplifier can be designed by means of designing the loop behavior. This design is not orthogonal to the design of weak distortion as both have to do with the design of the overall loop; for both, distortion minimization and bandwidth maximization, increasing the loop gain is favorable and therefore these optimizations are in the same direction and do not conflict.

Besides overall-loop measures, local-loop measures may also be helpful in the struggle for designing the frequency behavior without considerably degrading the distortion performance [11] (also see the previous section).

4.6.1 Frequency compensation

As stated earlier, a scaling factor is in principle frequency independent. For instance, in the (arbitrarily chosen) differential equation:

$$\alpha_1 \frac{de_o}{dt} + \alpha_0 e_o = e_i, \tag{4.8}$$

the scaling factors α_1 and α_0 are constant. Were they to have finite speed (a time dependence), the order of the differential equation would increase. Of course, this does not mean that the speed of practical scalers has to be infinite, or at least equal to the speed of light, their speed should be related to the maximum speed of the signals to be processed.

Therefore, a scaling factor is also described by a differential equation, for instance:

$$\alpha_2 \frac{d^2v}{dt^2} + \alpha_1 \frac{dv}{dt} + v = R_0 i, \tag{4.9}$$

where R_0 is the resistance for relatively slow signals and α_2 and α_1 are two coefficients. Designing the dynamic behavior of a circuit is not easily done by directly manipulating the differential equations. The use of the commonly known Laplace transform results in a more convenient model:

$$Z(s) = \frac{V}{I} = \frac{R_0}{\alpha_2 s^2 + \alpha_1 s + 1}, \tag{4.10}$$

where $Z(s)$ is the impedance as a function of the Laplace variable s, and V and I are the complex voltage and current, respectively. With this description the steady state response from the current, I, to the voltage, V, is found. The poles of this impedance have to be located in the s-plane such that the required frequency behavior is obtained.

The placement of the poles of the transfer between the input and output, i.e. the system poles, is the principal concern of frequency compensation, thus:

> for a given transfer, $H(s)$, in the complex domain described by:
>
> $$H(s) = \frac{H(0)}{s^2 + \alpha_1 s + \alpha_0}, \qquad (4.11)$$
>
> the frequency behavior is realized by giving the constants α_1 and α_0 the appropriate values.

This frequency compensation can be a tedious job. In order to reduce the number of design iterations and design time, design rules must help the designer at a relatively early stage, to tell him/her whether his/her design can succeed or not. Also, the type of model used for the amplifier may simplify the frequency compensation. Two types of methods can be distinguished:

- using the direct relation between frequency-compensation components and the characteristic polynomial of the system;
- using an intermediate step between the frequency compensation and the system poles.

An example of the first method is the one as described in [12] which is based on Cramer's rule as described in [13]. The method visualizes a circuit as an N-port, where N is the number of capacitors present in the circuit, and the corresponding ports are the terminals between which the capacitors are connected, see figure 4.3. The characteristic polynomial of the circuit is found from relatively simple calculations of DC port impedances under various conditions, i.e. a certain number of other ports are shorted.

This method is more suited for analysis purposes. It gives no insight into where compensation components have to be added. The correct compensation place has to be found by means of an exhaustive search. This is not permissible for short designing times.

The other method uses an intermediate step in the process of frequency compensation. The most commonly used is the root-locus method [14] which implicitly assumes feedback. This method uses the poles found in the loop, i.e. the *loop poles,* and the DC loop gain to find, by means of the construction rules of root loci, the actual *system poles.* As the loop poles are mostly related to explicit RC combinations and the construction rules for the root loci are

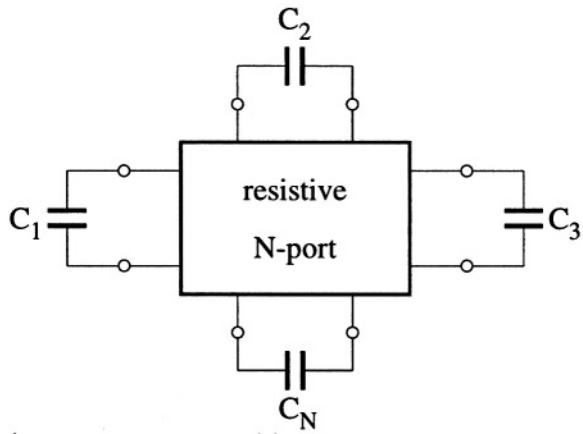

Figure 4.3: The circuit represented by a N-port with N external capacitors.

relatively simple, measures which have to be taken to end up with the desired system poles are relatively easily found.

The root-locus method fits very well on the asymptotic-gain model as shown in [3] and is therefore also used in this book. The frequency behavior of an amplifier can be split into two separate parts:

- absolute frequency behavior;
- relative frequency behavior.

This distinction is comparable with the distinction that is made when using polar coordinates. The absolute frequency behavior is proportional to the distance between the poles and the origin, i.e. the length of the place vector, and the relative frequency behavior has to do with the final relative pole positions, i.e. the angle between the place vector and the negative real axis. The absolute frequency behavior is explicitly determined by the speed capability of the constituent devices. Therefore, the design of the frequency behavior is split into two steps. First, the absolute frequency behavior has to be derived and made large enough and second, the loop poles have to be moved so that the system poles are at the desired relative positions in the s-plane. The *bandwidth* of a system is closely related to the absolute frequency behavior when the poles *have* the desired relative position. For instance, when the relative frequency behavior is of the Butterworth type, the absolute frequency behavior equals the bandwidth of the transfer. In the remaining discussion, the term bandwidth will be used for the absolute frequency behavior, remembering that for the final transfer the relative positions also have to be realized.

4.6.2 The maximal bandwidth

Designing the frequency behavior of an amplifier can be a lengthy job. When a designer has to conclude, after a lot of frequency-compensation trials, that the bandwidth capability of its amplifier is not high enough to reach the requirements, a lot of time and money is wasted. The LP product [3], which can be seen as a generalized GB product, is a measure of the maximum attainable bandwidth of an n-th order system.

4.6.2.1 The LP product

In chapter 2 the asymptotic-gain model was shown to be the appropriate model for the synthesis of amplifiers. Now the frequency dependency of the constituent elements is considered. The expression for the asymptotic-gain model then reads:

$$A_t(s) = A(s)_{t\infty}\frac{-A(s)\beta(s)}{1 - A(s)\beta(s)}, \tag{4.12}$$

in which the direct transfer ρ is ignored. This direct transfer cannot influence the pole positions; it can only introduce zeros into the system transfer. In expression (4.12) the asymptotic gain and the feedback factor are allowed to have some frequency dependency. This frequency dependence is a result of phantom zeros as discussed later on.

Assume a loop gain with n poles, $p_{li} = a_i + b_i j$ with $a_i < 0$, as given by:

$$A(s)\beta(s) = \frac{A(0)\beta(0)}{\prod_{i=1}^{n}(1 - s/p_{li})}, \tag{4.13}$$

where $A(0)\beta(0)$ is the DC loop gain. Then (a part of) the characteristic polynomial, CP, of $A_t(s)$ is given by:

$$CP(s) = s^n + \cdots + [1 - A(0)\beta(0)]\prod_{i=1}^{n}|p_{li}|. \tag{4.14}$$

The zero'th-order term is called the Loop-gain-Poles product, or LP product for short [3]. A more precise name would be the DC-return-difference-poles product, because the term $[1 - A(0)\beta(0)]$ is the return difference as defined in [15]. However, for accurate amplifiers, the magnitude of the loop gain is relatively large and the magnitude of the DC loop gain is approximately equal to the magnitude of the DC return difference. Expression (4.14) is found from the viewpoint of the root-locus method. However, of ultimate interest are the system poles. A part of the characteristic polynomial derived from the n system poles, $p_{si} = c_i + d_i j$ with $c_i < 0$, equals:

$$CP(s) = s^n + \cdots + \prod_{i=1}^{n}|p_{si}|. \tag{4.15}$$

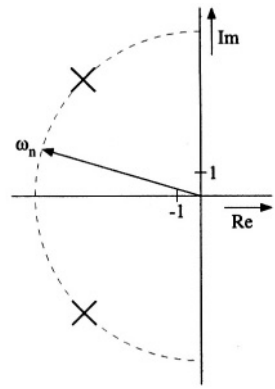

Figure 4.4: The characteristic place of Butterworth poles.

Here the zero'th-order term is the product of the moduli of all the system poles, i.e. the product of the lengths of the corresponding place vectors. Thus this term explicitly describes the absolute frequency behavior of the system. Consequently, the zero'th-order term found in equation (4.14) is a measure of the maximum attainable bandwidth of the corresponding system; for a first-order system, the LP product reduces to the GB product.

For amplifiers the Butterworth characteristic is a commonly used relative frequency behavior because it results in a maximum-flat-magnitude transfer. Therefore, in the rest of this chapter it will be assumed that a Butterworth characteristic is required unless explicitly stated otherwise.

For a Butterworth characteristic, the system poles are regularly placed on a half circle in the left half of the s-plane, see figure 4.4. For an nth-order system, the half circle is divided into n equal parts and in the middle of each part a pole is located. For a bandwidth of ω_n, the radius of the circle equals ω_n and thus the modulus of each pole equals ω_n. Applying this to equation (4.15) yields:

$$CP(s) = s^n + \cdots + \omega_n^n. \tag{4.16}$$

Comparing equation (4.14) and (4.16) yields the following relation that has to hold:

$$[1 - A(0)\beta(0)] \cdot \prod_{i=1}^{n} |p_{li}| = \omega_n^n. \tag{4.17}$$

The question is, which poles must be used to calculate this LP product?

> Example: What maximum bandwidth can be expected when the loop consists of three poles, $p_{l1} = -1$ kHz, $p_{l2} = -10$ kHz and $p_{l3} = -1$ GHz, and the DC loop gain equals -100?

When the three poles are used to calculate the LP product, the maximum bandwidth , B_{s3}, is found to be:

$$B_{s3} = \sqrt[3]{101 \cdot 1\ \text{kHz} \cdot 10\ \text{kHz} \cdot 1\ \text{GHz}} \approx 1\ \text{MHz}. \tag{4.18}$$

With a bit of experience one knows that the pole at -1 GHz is not dominant, i.e. it does not contribute to the bandwidth. The maximum bandwidth, B_{s2}, calculated on basis of p_{l1} and p_{l2} yields:

$$B_{s2} = \sqrt{101 \cdot 1\ \text{kHz} \cdot 10\ \text{kHz}} \approx 32\ \text{kHz}, \tag{4.19}$$

which is about a factor 30 lower than B_{s3}.

That the 1 GHz pole does not contribute to the bandwidth is clear; however, what to do when it was at -1 MHz. As was stated, the LP product only predicts the bandwidth when the poles used can be moved into the required relative positions, for this case the Butterworth positions; these poles then contribute to the bandwidth and are therefore called the *dominant poles.* Thus, only dominant poles should be used to calculate the LP product.

4.6.2.2 Dominant poles

In principle, one can see only at the end of the design what the maximum attainable bandwidth is. The LP product gives a maximum of that bandwidth at a earlier state. It is not certain whether or not this bandwidth can be reached. The only thing that is sure is that for the given number of stages the bandwidth cannot be larger than that indicated by the LP product. An analogous rule can be found for the dominant poles. The following derivation of the dominant poles is not limited to Butterworth behavior, it is generally applicable to other relative frequency behaviors as well. In contrast, the derivation of the dominant poles is limited to loops with only real poles which will be explained at the end of this section.

To find the dominant poles, the frequency behavior of the system is described again from two viewpoints. First, the characteristic polynomial is described from the loop viewpoint, which yields:

$$CP(s) = s^n - s^{n-1} \sum_{i=1}^{n} p_{li} + \cdots, \tag{4.20}$$

with $p_{li} < 0$. Second, the behavior is described as a function of the system poles, which yields:

$$CP(s) = s^n - s^{n-1} \sum_{i=1}^{n} p_{si} + \cdots, \tag{4.21}$$

with $p_{si} = c_i + d_i j$, $c_i < 0$, the system poles. Now the factor of the (n-1)th-order term is of interest. Comparing the term of equation (4.20) with the corresponding term of equation (4.21) yields:

$$\sum_{i=1}^{n} p_{li} = \sum_{i=1}^{n} p_{si}, \tag{4.22}$$

which states that the sum of the loop poles is equal to the sum of the system poles. From this property a criterion can be derived for the dominant poles.

The LP product gives a measure of the attainable bandwidth. As the required relative frequency behavior is known, the position of the system poles can be determined and from that their *sum* can be calculated. The *sum* of the loop poles is also given. These sums are generally not equal and frequency compensation has to be used as discussed in the next sections. All the methods discussed have the property of making the sum of the system poles smaller (i.e. more negative, remembering that the poles are negative). Thus, when the sum of the loop poles is smaller than the sum of the required system poles, frequency compensation will *not succeed*; the loop poles cannot be placed in the desired position; at least one loop pole is too far away from the origin. Such a pole will be called a non-dominant pole. The most negative pole from the loop has to be ignored and the LP product and the sum of the remaining poles has to be calculated again, et cetera, until the highest order of dominant poles is found. Thus:

> When p_{li} are the poles of the loop and p_{si} are the poles of the system, the dominant poles are the largest set of poles for which it holds that:
>
> $$\sum_{i=1}^{n} p_{li} \geq \sum_{i=1}^{n} p_{si}. \tag{4.23}$$
>
> The sum of the loop poles has to be less negative than the sum of the system poles.

Fulfilling this criterion is necessary but not sufficient. The characteristic polynomial may include more terms which must be given the appropriate values and it must be possible to implement the required frequency compensations in the circuit. In contrast, when the criterion is not fulfilled, it is certain that frequency compensation will not succeed with the given set of poles and loop gain, and the LP product of the set of dominant poles has to be increased. In [16] methods are described for systematically increasing the LP product of an amplifier.

Example:

> For the previous example the LP product for the third-order system predicted a bandwidth of 1 MHz. For a 1 MHz third-order Butterworth system, the sum of the poles equals -2 MHz ($p_{s1} = -1$

MHz, $p_{s2,3} = -\frac{1}{2}$ MHz $\pm \frac{1}{2}\sqrt{3}j$ MHz). The sum of the loop poles is approximately -1 GHz which is much smaller than -1 MHz and therefore at least p_{l3} is non-dominant. The predicted bandwidth of the second-order system is 32 kHz. The sum of the loop poles equals -11 kHz which is greater than the sum obtained from the system poles, -45 kHz. Thus the system has two dominant poles.

At the beginning of this section, the constraint was proposed that all the loop poles have to be real. This is required because for complex poles the contribution to the LP product can be relatively large while the contribution to the sum of the poles is relatively small as only the real parts count, the imaginary parts cancel. Complex poles can arise due to:

- LC-resonators[1];
- local feedback loops.

When these complex poles are non-dominant for the overall loop, they still can be dominant for the specific local loop and can even end up in the right-half plane. These poles should be properly damped by intervening in the corresponding local loop. Taking measures in the overall loop is likely to have only a slight effect since, at the frequencies for which the loop gain of the overall loop is reduced to one, the local loop may still have a considerable loop gain and thus the overall loop can no longer control the local loop.

When the complex poles are dominant, these poles have to be taken into account in the frequency compensation. Either the local loop should be broken in order to end up with real poles to be able to use the LP product and the dominant-pole criterion again, or the Rosenstark method [12] has to be used in which the system poles are directly manipulated; it does not use the notion of feedback and therefore it can be used for very complex networks. However, as was stated, it is based on analysis instead of synthesis.

A totally different strategy is to accept the number of loops and use techniques for synthesizing the frequency behavior of multi-loop amplifiers. As these techniques are not well established and they are beyond the scope of this book, they are not discussed here.

Further, the dominant-pole criterion is derived with the assumption that only compensation techniques are available which reduce the sum of the system poles. However, techniques also exist for increasing the sum of the system poles. These techniques use either negative feedforward, resulting in zeros in the right half plane, or positive feedback. These techniques have the property of reducing stability and are therefore less favorable and not used here.

[1]The resonator in itself can also be seen as a loop, in which energy is going round from L to C to L, et cetera.

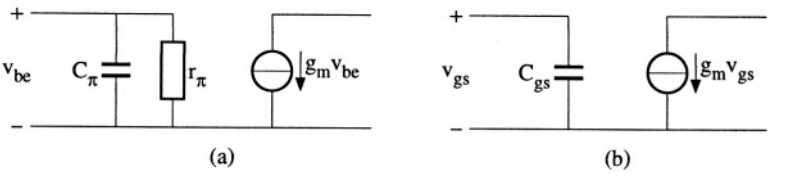

Figure 4.5: The relevant model for active devices. a) The bipolar transistor and b) the field-effect transistor.

4.6.3 Moving around the poles

In the previous sections the maximum attainable bandwidth of an amplifier was found. Nothing was said about *how* to reach this bandwidth nor the *possibility* of reaching it. The only thing that can be said is that when the LP product is too low, it is not possible to reach the required bandwidth at all.

This section discusses the *placement* of the poles, i.e. obtaining the required relative frequency behavior. The frequency compensation techniques are not allowed to interfere with the earlier design steps, or at least only up to a negligible level. Thus any additional noise contribution and distortion must be kept within acceptable levels and the reduction of the LP product due to the frequency compensation must be as low as acceptable, ideally no reduction.

To ease the frequency compensation any further, the small-signal diagram for the active devices is limited to only the relevant part [16], see figure 4.5. After the frequency compensation using these simple models, the models can be gradually extended with their parasitics. When a parasitic has a non-negligible influence, its influence can be reduced, generally to a large extent, by means of additional design steps which are orthogonal to the former steps [16], for instance, adding current buffers in order to reduce the Miller effect. When the influence of the parasitic cannot be reduced as much as required, one iteration of the frequency compensation has to be done taking this parasitic into account.

In this book frequency compensation is assumed to be the addition of *passive networks* to a circuit in order to alter the position of system poles. The most simple situation occurs in the case of two poles[2]. Figure 4.6 depicts a typical root locus of a non-compensated amplifier. Clearly, the sum of the loop poles is too high for obtaining system poles which are in the Butterworth position. Consequently, the system poles become relatively complex. To obtain Butterworth behavior, the sum of the poles has to be reduced. Four methods are possible for achieving this, see figure 4.7. Figure 4.7a depicts the situation in which the real part of *one* pole is reduced. As the compensation networks are passive and thus cannot increase the LP product, this action inherently reduces

[2]Frequency compensation of a first-order system is not necessary as it reaches the bandwidth given by the LP product without compensation techniques.

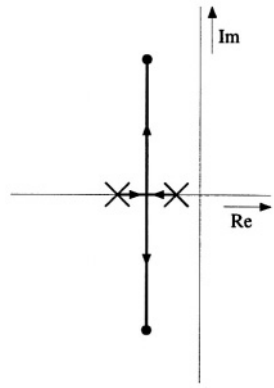

Figure 4.6: A typical root locus of a non-compensated amplifier.

the DC loop gain. An example of this method is resistive broadbanding.

Figure 4.7b depicts the situation in which *two* poles are split. One pole is shifted towards the origin, which is done by introducing for relatively low frequencies an extra attenuation in the loop of the amplifier. For frequencies beyond the second original pole, this attenuation is gradually relieved, resulting in a zero canceling this second pole and finally, when attenuation is completely canceled, a new pole is found. This pole is a factor δ away from the original second pole, which equals the factor by which the original first pole was shifted to the origin. Thus the sum of the poles is reduced and the LP product is not degraded. An example of this method is pole-zero cancellation.

Figure 4.7c depicts pole splitting which introduces interaction between the *two* poles such that they split. No intermediate zero is used. An example is the technique called pole-splitting.

Figure 4.7d shows the use of a zero to bend the root locus. In contrast to the earlier techniques, this method alters the position of the system poles by influencing the root locus without altering the position of the loop poles. In order to obtain an all-pole system transfer this zero has to be a phantom zero. Further, the techniques depicted in figure 4.7a,b and c only influence at most two poles, whereas the phantom zero technique of figure 4.7d can exert an influence on all the poles.

For higher-order systems a combination of these techniques can be applied. Generally, for an n-th order system, n-1 frequency compensations are required.

In the following sections the four techniques are discussed and the influence on the LP product is derived.

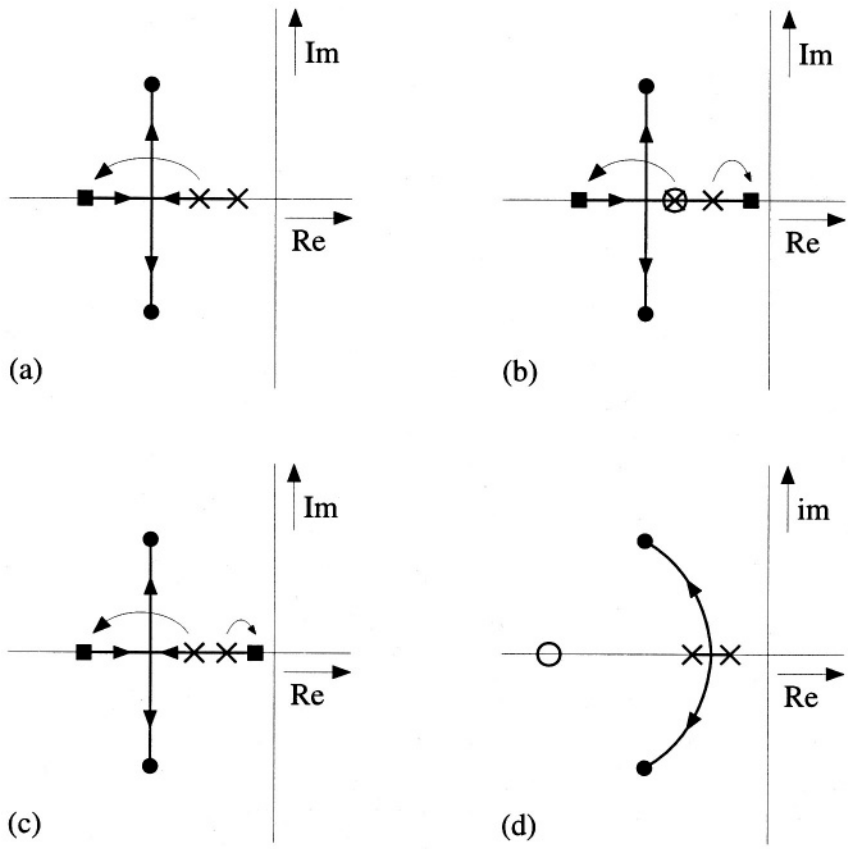

Figure 4.7: Four methods to alter the pole pattern of an amplifier.

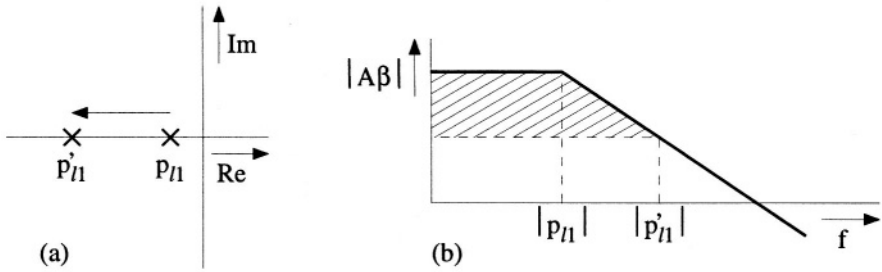

Figure 4.8: The basic idea of resistive broadbanding. The influence on a) the pole-zero plot and b) the loop gain.

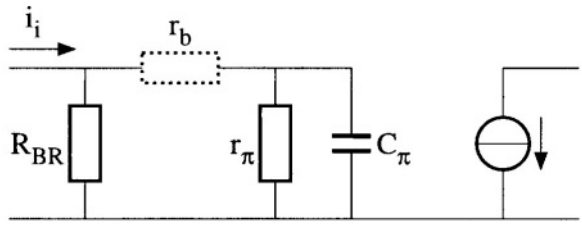

Figure 4.9: A passive implementation of resistive broadbanding.

4.6.3.1 Resistive broadbanding

Resistive broadbanding acts on one pole only. The basic idea of resistive broadbanding is depicted in figure 4.8. With a compensation network a single pole is shifted further from the origin. The factor by which the DC loop gain reduces is equal to the factor by which the pole has shifted downwards.

Resistive broadbanding can be realized in two ways, passive and active. Figure 4.9 shows a passive implementation. The original pole shifts a factor $1 + r_\pi/R_{BR}$ downwards, the DC loop gain is reduced by the same factor and thus the LP product remains the same. Adding base resistances to the model reduces the LP product by a factor:

$$\frac{LP_{before}}{LP_{after}} = 1 + \frac{r_b}{R_{BR}}, \tag{4.24}$$

where r_b is the base resistance. In the original case, i.e. without R_{BR}, for relatively high frequencies the complete input current, i_i, flows through r_π and c_π. In the case of a finite R_{BR}, for relatively high frequencies input current i_i divides over R_{BR} and r_b giving a reduction of the gain the stage contributes to the overall amplifier loop gain. This current division can be removed by adding an inductor in series with R_{BR} [3], resulting in the original LP product.

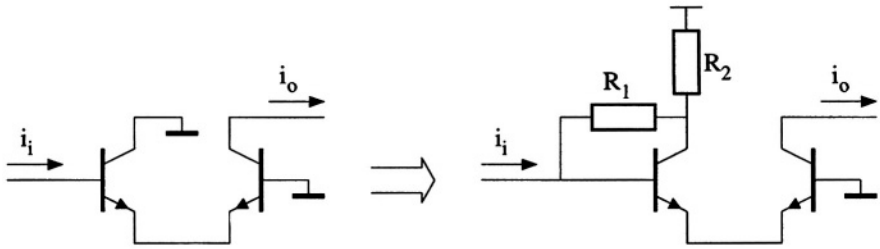

Figure 4.10: Resistive broadbanding by means of local feedback.

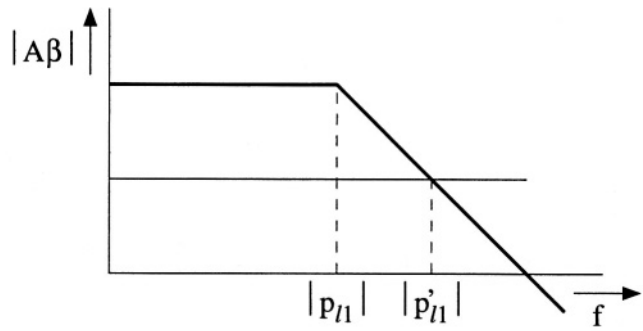

Figure 4.11: The influence of resistive broadbanding, by means of a local loop, on the bode plot.

This method of implementing resistive broadbanding has two drawbacks. First, the gain by which the overall gain is reduced, is totally wasted; this is indicated by the dashed area in figure 4.8. Nothing is done with it, resulting in an increased distortion level. Second, for relatively low values of R_{BR} the LP product reduces considerably.

Resistive broadbanding by means of local feedback does not have these drawbacks. An example is given in figure 4.10. By means of the current-feedback network the transfer of the differential pair is reduced to:

$$\frac{i_o}{i_i} = 1 + \frac{R_1}{R_2}. \tag{4.25}$$

The method is elucidated by the bode plot in figure 4.11. The thick line indicates the original transfer and the thin line is the ideal transfer. At the intersection point of these two lines, the loop gain is 1 and the new pole is found, p'_{l1}. Now the loop gain reduction is not wasted but is used in a local feedback, this local stage is linearized. However, as the total loop gain is reduced, the non-linearities of other stage are less suppressed, resulting in a slight increase of the distortion.

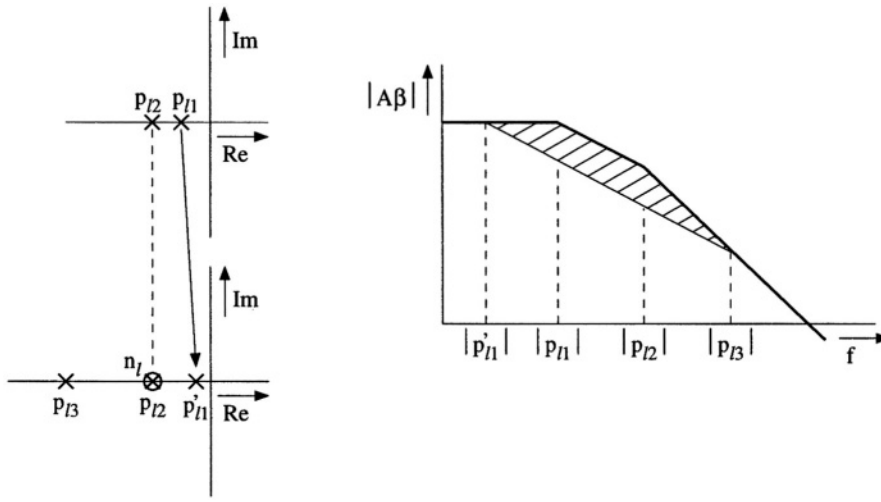

Figure 4.12: The principle of pole-zero cancellation.

The new pole position is found at approximately:

$$p'_{l1} \approx p_{l1} \cdot \frac{R_2}{R_1 + R_2}. \tag{4.26}$$

The big difference with the previous type of implementation is that the impedance level of the feedback network can freely be chosen, up to a certain extent, of course. Prom exact calculations the following reduction factor of the LP is found:

$$\frac{LP_{before}}{LP_{after}} \approx 1 + \frac{2r_b}{R_1 + R_2}, \tag{4.27}$$

where r_b is the base resistance of one transistor and, assuming that $R_2 \gg 1/g_m$, $r_b \ll (R_1 + R_2)$ and the DC loop gain of the local loop is much larger than one. The LP-product reduction is now caused by the remaining high frequency current division between the input impedance of the differential pair, $2r_b$, and the series connection of the two feedback resistors, $R_1 + R_2$.

4.6.3.2 Pole-zero cancellation

Pole-zero cancellation is a method for splitting two poles, i.e. the sum of the poles reduces. The principle is depicted in figure 4.12. One pole is shifted towards the origin; as a result a zero can be created to cancel another pole and inherently, a new pole is found because the LP product cannot increase.

In figure 4.13 a straightforward implementation is given of pole-zero cancellation. With a capacitor a pole is shifted closer to the origin, p'_{l1}. When at

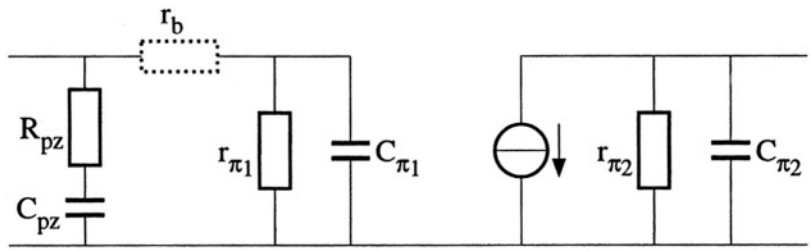

Figure 4.13: A straightforward implementation of pole-zero cancellation.

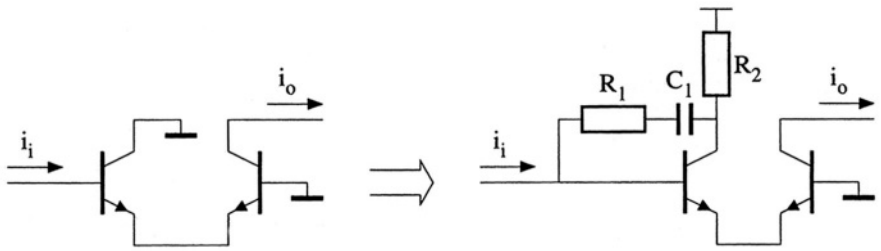

Figure 4.14: Pole-zero cancellation by means of local feedback.

higher frequencies the influence of this capacitor is removed again by a resistor, a zero is obtained, n_l. With this zero another pole, p_{l2} can be canceled. For even higher frequencies, R_{PZ} and c_π result in a new pole, p_{l3}. When calculating the two new poles, assuming that the zero cancels pole p_{l2}, it is easily found that the LP product does not change.

Again introducing the base-resistances, the loop gain reduces by a factor:

$$\frac{LP_{before}}{LP_{after}} = 1 + \frac{r_b}{R_{PZ}}. \tag{4.28}$$

By using the pole-zero cancellation in a local feedback configuration, the influence of the base resistance and the effect of reduced loop gain (cf. resistive broadbanding) is diminished. This principle is depicted in figure 4.14. This figure depicts a single-side driven and loaded differential pair. The pole-zero cancellation is implemented by means of R_1, R_2 and C_1 which realize a frequency-dependent current feedback. The influence on the Bode plot is depicted in figure 4.15. Originally, the current transfer of the differential pair equals the current-gain factor β_f, with a pole at f_T/β_f; the thick line in figure 4.15. The asymptotic gain of this stage including the local feedback is indicated by the thin line. At the intersection points of the thin and thick line the loop gain is again 1 and the actual poles of the new transfer are found. The zero in the asymptotic gain is at a frequency for which the loop gain is relatively high

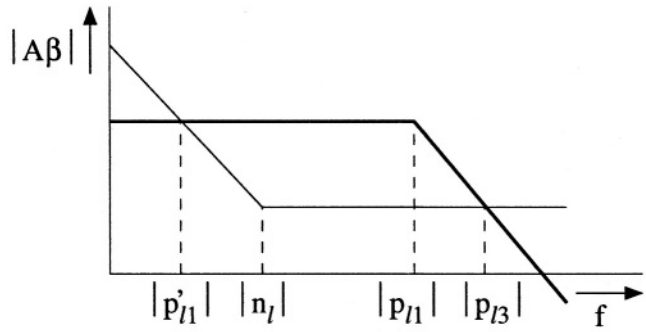

Figure 4.15: The influence of a local current feedback on the transfer of a differential pair.

and thus this zero is also found in the new transfer. It is given by:

$$n_l = \frac{-1}{2\pi C_1(R_1 + R_2)}. \tag{4.29}$$

With this zero a pole of another stage can be canceled.

The influence of the base-resistance for this type of pole-zero cancellation is also significantly reduced; the decrease in LP product is only:

$$\frac{LP_{before}}{LP_{after}} = 1 + \frac{2r_b}{R_1 + R_2}, \tag{4.30}$$

which is the same result as was found for resistive broadbanding implemented by means of local feedback. This is easily understood when it is noticed that for relatively high frequencies the two stages tend to the same equivalent circuit.

4.6.3.3 Pole splitting

Pole splitting is a technique which splits two poles by introducing an interaction between two poles by means of a local loop. The principle is depicted in figure 4.16. The poles are split apart while their product, ideally, remains constant such that the LP product is not changed. In figure 4.17 an example is given of pole-splitting. Capacitor C_{SP} acts as a Miller capacitance, the poles at the input and output are split by means of local feedback. The reduction of the LP product can readily be found to be equal to:

$$\frac{LP_{before}}{LP_{after}} = 1 + C_{SP}\left(\frac{1}{C_1} + \frac{1}{C_2}\right). \tag{4.31}$$

The more the poles are split, the lower the LP product. The level of splitting is determined by C_{SP} and by the voltage gain between the two nodes between

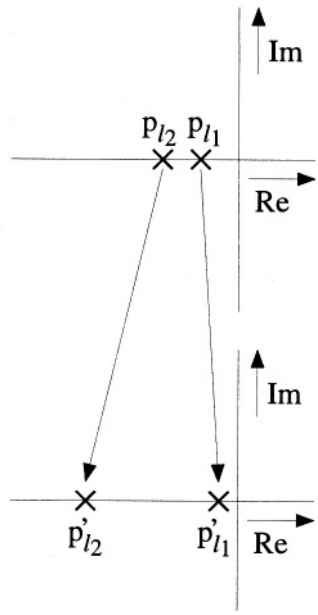

Figure 4.16: The principle of pole splitting.

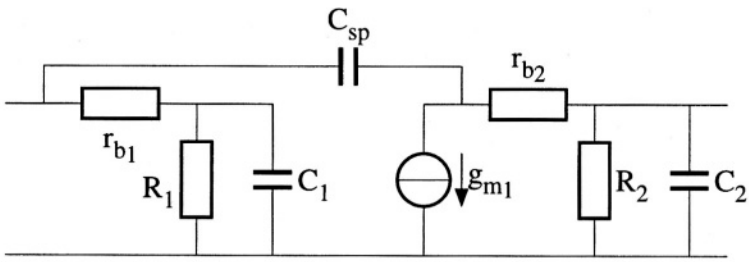

Figure 4.17: Pole splitting in a circuit.

which C_{SP} is connected. For higher voltage gains, C_{SP} can be smaller in order to end up with the same amount of splitting, and thus less LP product is lost. Therefore, stages with a high voltage gain are the best stages to introduce this type of pole splitting.

Introducing the base resistances in the circuits results in the following approximated expression for the LP-product reduction:

$$\frac{LP_{before}}{LP_{after}} = 1 + C_{SP}\left(\frac{1}{C_1} + \frac{1}{C_2}\right) + g_{m1}r_{b2}\frac{C_{SP}}{C_1}, \tag{4.32}$$

in which it is assumed that the driving impedance for the input is negligibly large, and the base resistances are relatively small compared to R_1 and R_2. Clearly, the reduction due to the base resistances can be ignored for lower g_ms.

Compared to pole-zero cancellation, this method requires less capacitance to achieve the same splitting as the voltage gain of a stage is used. However, pole-splitting by means of interaction costs more LP product. Further, due to C_{SP} a right-half plane zero is introduced. The stage is no longer unilateral which can be a severe problem for the stability.

There are several methods for reducing the effect of this right-half plane zero. Reference [17] describes the use of a voltage follower in order to obtain a unilateral stage; the zero is *removed.* In [11] a series resistor is used to *compensate* for the zero. This resistor has to be equal to $1/g_m$. But, as the collector current of the transistor is not constant as a result of an applied signal, g_m varies and perfect compensation of the zero is not achieved. The resulting pole-zero doublet is disadvantageous for the settling time. Reference [18] summarizes the different active buffering techniques. In [19] a different method is introduced. Here multi-path techniques are used to *remove* the right-half plane zero.

The active buffer techniques and the multi-path techniques completely remove the zero. The technique using the resistor compensation does not. The effect of the compensation technique is studied here in more detail. With the additional resistor, R_{PZ}, see figure 4.18, the zero is found at:

$$n_l = \frac{+1}{2\pi C_{SP}(\frac{1}{g_{m1}} - R_{SP})}. \tag{4.33}$$

For the zero there are three possibilities:

- $R_{SP} < \frac{1}{g_{m1}}$: zero in RHP;
- $R_{SP} = \frac{1}{g_{m1}}$: zero at infinity;
- $R_{SP} > \frac{1}{g_{m1}}$: zero in LHP.

This third case seems to be the most advantageous, pole-spliting ánd a LHP zero. However, when calculating the characteristic polynomial, a third pole is

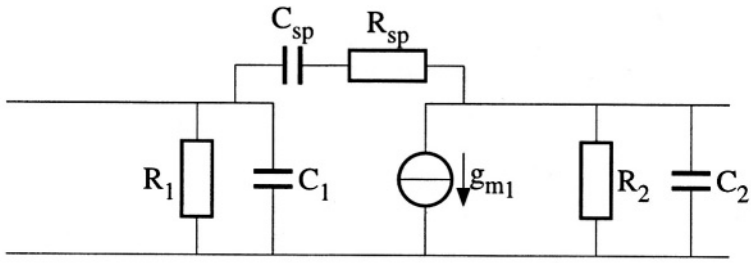

Figure 4.18: The compensation of the right-half plane zero by a resistor.

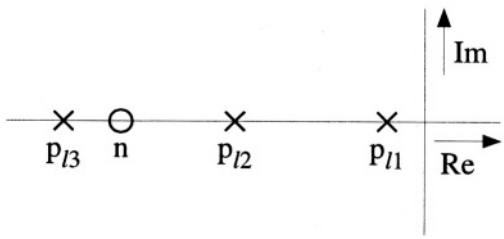

Figure 4.19: The pole-zero plot of pole splitting with $R_{SP} > \frac{1}{g_{m1}}$.

found at:

$$p_{l3} = \frac{-1}{2\pi R_{SP} C_{SP}} \cdot \left[1 + C_{SP}\left(\frac{1}{C_1} + \frac{1}{C_2}\right)\right], \tag{4.34}$$

with the last factor equal to the factor of equation (4.31). Thus for relatively small split capacitors, the pole is found at:

$$p_{l3} = \frac{-1}{2\pi R_{SP} C_{SP}}. \tag{4.35}$$

With exact compensation, $g_{m1} = R_{SP}$, the zero in the RHP vanishes, but an additional pole is found at the same position in the LHP. The additional loop gain due to the zero is removed but the additional phase shift remains. For the third situation the pole-zero pattern of figure 4.19 applies. The zero is now found in the LHP closer to the origin than the third pole. The fraction between this pole and zero equals the factor by which the LP product is reduced, equation (4.31). For relatively small split capacitors, the pole and zero cancel each other. The pole and zero are a reasonable factor apart only when LP product is lost due to a relatively large amount of splitting.

4.6.3.4 Phantom zeros

Phantom zeros [14] are zeros which are realized in the feedback factor β [3]. They are thus realized *outside* the nullor implementation. In the case of a zero

in β, the relevant part of the asymptotic-gain model is given by:

$$A_t(s) = \frac{1}{\beta(0)} \cdot \frac{-A(s)\beta(0)}{1 - A(s)\beta(0)(1 - s/n)}, \tag{4.36}$$

where n is the zero. As can be seen, the zero in the denominator of the asymptotic-gain part cancels with the zero in the numerator of the second factor. The zero is only effectively found in the denominator of the second factor and can therefore be used for the frequency compensation. The characteristic polynomial of a second-order system when one phantom zero is introduced is given by:

$$CP(s) = s^2 - s\left[p_{l1} + p_{l2} - \frac{A(0)\beta(0)p_{l1}p_{l2}}{n_l}\right] + p_{l1}p_{l2}[1 - A(0)\beta(0)]. \tag{4.37}$$

As can be seen, the LP product does not change as a result of the phantom zero. Of course, when a phantom zero is practically realized, influences via base resistances, and so on, may also occur. However, the phantom zero is generally near the band edge of a system and therefore the resulting second-order effects will be far beyond the band edge. For an n-th order system, (n-1) phantom zeros are required to alter the sum of the system poles.

A phantom zero is realized when an attenuation in the feedback network is removed beyond a certain frequency. The effectiveness of the phantom zero is determined by the level of the attenuation that is removed. The higher this attenuation is, the more effective this phantom zero is. This can be seen when the unavoidable accompanying pole is examined. Assume that in β a reduction of a factor δ is removed beyond a frequency corresponding to n. Then the accompanying pole is given by:

$$p = n \cdot \delta. \tag{4.38}$$

This is in the case of a single phantom zero; the reduction is removed by means of a first-order behavior.

An example is given in figure 4.20. Originally, the current from the feedback resistor, R_{fb}, was divided between C_s and C_i. This resulted in a reduction of $(1 + C_s/C_i) = \delta$. With resistor R_{ph}, the current path via C_s is made less favorable with respect to the current path via C_i beyond the frequency f_{ph}:

$$f_{ph} = \frac{1}{2\pi R_{ph} C_s}. \tag{4.39}$$

The accompanying pole is found at:

$$p = \frac{-1}{2\pi R_{ph} \frac{C_s C_i}{C_s + C_i}}. \tag{4.40}$$

This pole is a factor δ away from the phantom zero.

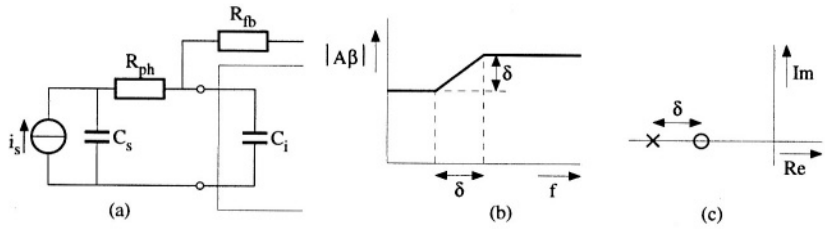

Figure 4.20: An example of the effectiveness of a phantom zero. a) The realization of a phantom zero by R_{ph}, b) the influence on the magnitude of the loop gain and c) the pole-zero plot.

4.7 Conclusion

In this chapter the structured design of amplifiers was discussed. It was shown that the three fundamental design aspects: noise, distortion and bandwidth can be treated orthogonally in the design process.

Noise is mainly determined at the input stage and depends on the type of source and input device. Once minimized, the noise performance of the amplifier can no longer improve.

Distortion was divided into clipping distortion and weak distortion. Clipping distortion is mainly found at the output as the signals are the largest there. Weak distortion can be caused by the active devices by means of g_m and β distortion, of which g_m distortion is the most severe one. It has been shown that local feedback makes no sense as a measure for reducing distortion. The best way to reduce distortion is by increasing the overall loop gain.

The design of the bandwidth is shown to be governed by the overall loop. Thus bandwidth capability can be improved everywhere in the loop. The LP product is a measure of the maximum attainable bandwidth, i.e. the absolute frequency behavior. This requires the identification of the dominant poles. The dominant poles can be derived from the notion that the sum of the system poles and the sum of the loop poles remains constant when closing the loop. After realizing a sufficient LP product, the poles have to be moved to end up with the required relative frequency behavior, for instance Butterworth.

Four types of frequency compensation methods have been discussed and their influence on the LP product, i.e. the bandwidth capability, was investigated. For the sake of distortion performance, frequency compensation techniques using local feedback are favorable as they use the portion of loop gain, by which the overall loop gain is reduced, for linearizing the stage which is locally fed back. Techniques not using local feedback completely waste this portion.

Bibliography

[1] J.E. Solomon. The monolithic op amp: A tutorial study. *IEEE Journal of Solid-State Circuits,* 9(6):314–332, December 1974.

[2] P.R. Gray and R.G. Meyer. MOS operational amplifier design - a tutorial overview. *IEEE Journal of Solid-State Circuits,* 17(6):969–982, December 1982.

[3] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[4] D.R. Frey. Exponential state space filters: A generic current mode design strategy. *IEEE Transactions on Circuits and Systems I*, 43(l):34–42, January 1996.

[5] Z.Y. Chang and W.M.C. Sansen. *Low-Noise Wide-Band Amplifiers in Bipolar and CMOS Technologies.* Kluwer Academic Publishers, Dordrecht, 1991.

[6] J. Davidse. *Analog Electronic Circuit Design.* Prentice Hall International (UK) Ltd, London, 1991.

[7] P.R. Gray and R.G. Meyer. *Analysis and Design of Analog Integrated Circuits.* John Wiley & Sons Inc., New York, 1993.

[8] H.T. Friis. Noise figures of radio receivers. *Proceedings I.R.E.,* 32:419–422, 1944.

[9] MicroSim Corporation. *Manual Pspice 4.05.*

[10] E.M. Cherry and D.E. Hooper. *Amplifying Devices and Low-Pass Amplifier Design.* John Wiley and Sons, New York, 1968.

[11] E.M. Cherry. A new result in negative-feedback theory, and its application to audio power amplifiers. *IEEE Journal on Circuit Theory and Applications,* 6(3):265–288, July 1978.

[12] S. Rosenstark. Re-examination of frequency response calculations for feedback amplifiers. *International Journal of Electronics,* 58(2):271–282, 1985.

[13] B.L. Cochrun and A. Grabel. A method for the determination of the transfer function of electronic circuits. *IEEE Transactions on Circuit Theory,* 20(1):16–20, January 1973.

[14] M.S. Ghausi and D.O. Pederson. A new design approach for feedback amplifiers. *IRE Transactions on Circuit Theory,* 8:274–284, 1961.

[15] H.W. Bode. *Network Analysis and Feedback Amplifier Design.* Van Nostrand, New York, 1945.

[16] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

[17] Y.P. Tsividis and P.R. Gray. An integrated NMOS operational amplifier with internal compensation. *IEEE Journal of Solid-State Circuits,* 11(6):748–753, December 1976.

[18] C.A. Makris and C. Toumazou. Current-mode active compensation techniques. *Electronics Letters,* 26(21):1792–1794, October 1990.

[19] R.G.H. Eschauzier, L.P.T. Kerklaan, and J.H. Huijsing. A 100-MHz 100-dB operational amplifier with mulitpath nested miller compensation structure. *IEEE Journal of Solid-State Circuits,* 27(12):1709–1716, December 1992.

Chapter 5

Harmonic oscillators

5.1 Introduction

Frequency references are widely-used building blocks. They are found in almost every electronic system, from the local oscillator in the down-conversion part of a communication instrument to the oscillator in wristwatches [1]. Oscillators used as a reference require a relatively high frequency stability as a function of time. For oscillators in communication instruments, short-term stability is very important, i.e. the noise power of the oscillator that is relatively close to the carrier. High short-term stability is required to prevent the receiver in a communication system, for instance, from mixing not only the desired channel to an intermediate frequency but also an adjacent channel. For the oscillator in a wristwatch this short-term stability is not of prime importance. It does not matter, for most applications, when the time of the day cannot be read to an accuracy of micro-seconds. For this oscillator a very low-frequency stability is important. When the frequency reference generates a signal at a frequency of 1 Hz, this frequency should not be 1.1 Hz a week later as the watch would then run too fast and it would be useless.

Independently of which type of stability is required, short-term stability, long-term stability, et cetera, all frequency references are similar in that the frequency of the signal they generate has to be well known and they have to be based on a phenomenon which has an intrinsic frequency of preference to guarantee high stability.

5.2 The basic function

Periodical solutions from differential equations can be used to implement a frequency reference. For this purpose, the amplitude of the solution should not

decay as a function of time as otherwise the reference signal becomes too small to be processed correctly. For the second-order linear homogeneous differential equation:

$$\frac{d^2e(t)}{dt^2} + \omega_0^2 e(t) = 0, \tag{5.1}$$

the following solution holds:

$$e(t) = A\sin(\omega_0 t + \phi_0), \tag{5.2}$$

where A and ϕ_0 are the two boundary conditions. This expression describes a sinusoidal signal with a constant frequency, ω_0, and amplitude, A. The frequency is determined by a constant coefficient of the differential equation and can therefore be used as a reference, in contrast to the constant amplitude which is given by an initial condition and is therefore subject to uncertainty.

In Chapter 2, this differential equation and solution were found to describe an ideal harmonic oscillator. The other frequency references which were found are inherently contaminated with noise due to the power dissipation in the real poles involved. The intrinsic noise performance of harmonic oscillators is high; ideally they can be noise free and thus by proper design it is possible to obtain an oscillator with a very low noise level and thus a high frequency stability.

The harmonic oscillator is therefore the correct choice for frequency references for which a low-noise performance is inevitable; in contrast to the first-order oscillator which has no frequency of preference and is consequently easily tuned by noise [2][1].

Components, i.e. resonators, which can be used as the core of a harmonic oscillator are the crystal and the LC tank, for instance. In principle, each phenomenon that has some kind of intrinsic frequency selectivity may be used. These components have a (simplified) behavior according to:

$$\frac{d^2e(t)}{dt^2} + 2\zeta\omega_0\frac{de(t)}{dt} + \omega_0^2 e(t) = 0, \tag{5.3}$$

where ζ models the damping in the components as a result of the (low) dissipation in the resonator. The solution of this differential equation is given by:

$$e(t) = A \cdot \exp(-\zeta\omega_0 t) \cdot \sin\left(\omega_0\sqrt{1-\zeta^2}\cdot t + \phi_0\right), \quad 1-\zeta^2 > 0. \tag{5.4}$$

Compared to equation (5.2), an additional factor comprising the damping, ζ, is found. For passive circuits the damping is always larger than zero and thus the amplitude of the vibration decays as a function of time. To get a signal with a constant amplitude, the damping term has to be nullified by an active block which compensates for the losses in the resonator, see figure 5.1. As the

[1]This property is profitably used when tunable oscillators are required with moderate noise specifications.

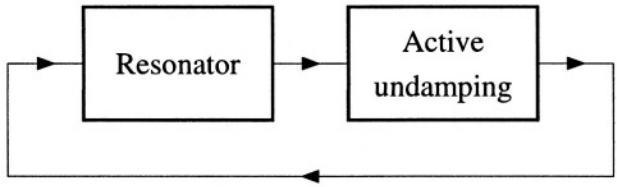

Figure 5.1: The compensation of the losses in a resonator by means of an active undamping.

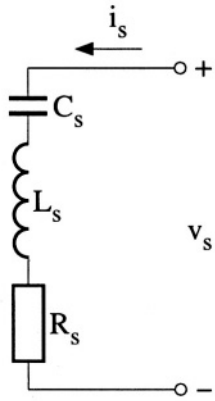

Figure 5.2: A series resonator.

undamping supplies power to the resonator, this adds noise to the resonator signal and degrades its performance. This should be kept to an acceptable level by means of proper circuit design.

5.2.1 The resonator

In this chapter the discussions concentrate on the use of an LC tank as the resonator in the frequency reference. The discussions for the crystal are more or less analogous and the crystal is therefore not treated explicitly.

Let us assume a series resonator as depicted in figure 5.2. The v-i relation of the series resonator is described by:

$$\frac{d^2 i_s(t)}{dt^2} + \frac{di_s(t)}{dt}\frac{R_s}{L_s} + i_s(t)\frac{1}{L_s C_s} = \frac{dv_s(t)}{dt}\frac{1}{L_s}, \tag{5.5}$$

where $i_s(t)$ and $v_s(t)$ are the resonator current and voltage as a function of time,

respectively. The homogeneous solution of this equation is given by:

$$i_s(t) = A \cdot \exp\left(-\frac{R_s}{2L_s}t\right) \cdot \sin\left[\sqrt{\frac{1}{L_sC_s} - \left(\frac{R_s}{2L_s}\right)^2} \cdot t + \phi_0\right]. \tag{5.6}$$

As can be seen, the damping only depends on the quality of the inductor [2], i.e. the ratio of R_s and L_s. Introducing the quality factor of the resonator [3], Q_s, into this expression, which is defined as (for $Q_s \gg 2\pi$):

$$Q_s = 2\pi \frac{\text{total energy stored}}{\text{energy lost per cycle}} = \frac{\omega_0 L_s}{R_s}, \tag{5.7}$$

where $\omega_0 = \frac{1}{\sqrt{L_sC_s}}$, yields:

$$i_s(t) = A \cdot \exp\left(-\frac{\omega_0}{2Q_s}t\right) \cdot \sin\left[\omega_0\sqrt{1 - \left(\frac{1}{4Q_s^2}\right)^2} \cdot t + \phi_0\right]. \tag{5.8}$$

From this equation it follows directly that for N cycles, a fraction δ is left of the original amplitude with N and δ related according to:

$$N = \frac{Q}{\pi} \ln \delta. \tag{5.9}$$

The resonator signal is depicted in figure 5.3 for $Q_s = 10$. From this plot and from equation (5.9) it is found that after about 2.2 cycles the amplitude is decreased to half the original value.

When the damping is nullified, by means of undamping or in the case of an ideal resonator, i.e. with an infinite Q_s, the frequency of oscillation equals:

$$\omega_o = \frac{1}{\sqrt{L_sC_s}}, \tag{5.10}$$

which can be used as a reference frequency.

5.2.2 The undamping

In figure 5.4 the series resonator is depicted with its undamping. To find the port relation (impedance), $Z_u(t) = v_s/(-i_s)$, of the undamping, equation (5.5) is extended to:

$$\frac{d^2i_s(t)}{dt^2} + \frac{di_s(t)}{dt}\frac{R_s}{L_s} + \frac{d[Z_u(t) \cdot i_s(t)]}{dt}\frac{1}{L_s} + i_s(t)\frac{1}{L_sC_s} = 0. \tag{5.11}$$

[2]It should be noted that the capacitor may also have some losses due to contact resistances and losses in the dielectric, for instance. It is assumed here that the inductor is the main cause of the losses.

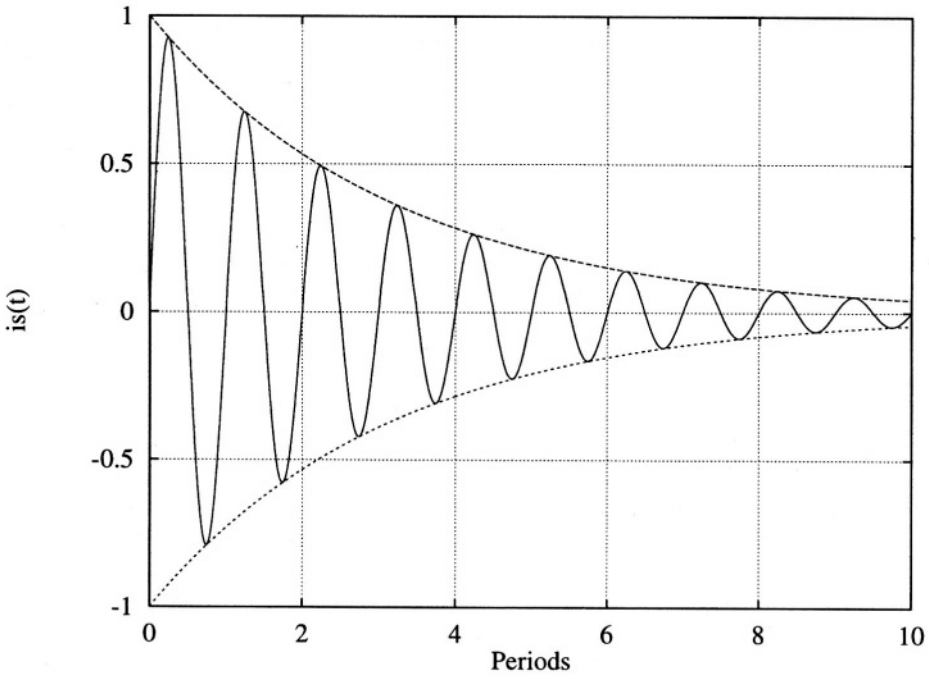

Figure 5.3: The output signal of a resonator as a function of time (normalized to the period time) for $Q_s = 10$; the dotted line indicates the envelope.

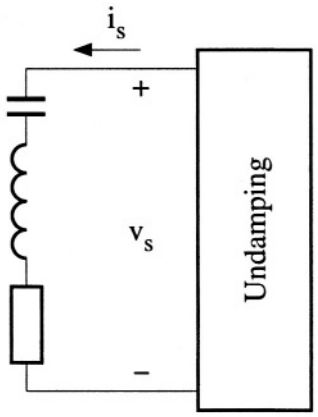

Figure 5.4: The series resonators with its undamping.

To obtain exact compensation for the losses, the following has to hold:

$$\frac{di_s(t)}{dt}\left[R_s + Z_u(t)\right] + \frac{dZ_u(t)}{dt} i_s(t) = 0. \tag{5.12}$$

Assuming for the moment that Z_u is constant, equation (5.12) simplifies to:

$$Z_u = -R_s. \tag{5.13}$$

The port relation of the undamping is described by a negative resistance whose magnitude equals the resistance of the inductor and thus the undamping *supplies* power to the resonator that is equal to the power losses in the inductor. Consequently, the resonator-undamping combination will oscillate with a constant amplitude.

For practical resonators the losses of the inductor are subject to statistical spread. For a resonator with losses higher than the expected losses, the undamping is not sufficient and the oscillator does not startup. Therefore, the initial undamping has to be chosen to be larger than the expected losses to ensure startup for most of the resonators. When a time and amplitude independent undamping is chosen, the output signal will have an exponentially increasing amplitude as it is over-undamped. Therefore, the undamping impedance has to be made amplitude dependent. Roughly speaking, two types can be distinguished:

- time-averaged control;
- instantaneous control.

For time-averaged control, the undamping impedance is controlled as a function of the difference between the mean amplitude of the oscillator signal and the required amplitude. As this control contains an integration (the determination of the mean value), it controls the amplitude to the required value. Speaking in terms of equation (5.12), for relatively long times, the variation in $Z_u(t)$ becomes smaller and smaller as it approaches the required value and therefore the second term in equation (5.12) can be ignored. After some time this time-averaged control sets the undamping impedance to $-R_s$.

In the case of instantaneous control, the value of the undamping impedance is related to the momentary value of the resonator signal. The magnitude of the impedance reduces for larger resonator signals. For equation (5.12) this means that the last term cannot be ignored as it has to compensate for the over undamping described by the first term. Noting that for each period the losses have to be compensated, the following relation has to be fulfilled:

$$\int_0^T i_s^2(t) R_s dt = \int_0^T i_s^2(t) Z_u(t) dt. \tag{5.14}$$

Of course, the time dependence of $Z_u(t)$ depends on the relation between the momentary value of the oscillator signal and $Z_u(t)$. In [4] it is shown that an over undamping by a factor of two leads to a degradation of the frequency stability by 3 dB due to noise folding caused by the intrinsic non-linearity of instantaneous control.

5.3 Relation to the fundamental design aspects

In Chapter 2, noise, signal power and bandwidth were found as the three fundamental design aspects of electronic circuits. An oscillator is used in a larger system and from that system the required behavior of the oscillator is found. The system requires, for instance, a frequency reference with a certain signal power and carrier-to-noise ratio. Further, specifications are found with respect to bandwidth, i.e. the noise floor for relatively high frequencies is not allowed to be beyond a specified value. When impedances and loading are taken into account, the value and bandwidth of the output impedance may be specified.

The performance with respect to these design aspects is bounded by the practical limits of the quality of the principal frequency reference, i.e. the resonator. For retrieving a signal from the oscillating resonator-undamping combination, the undamping is very often provided with an additional output to which to connect the load. The quality of the signal finally obtained at the output of the oscillator is therefore determined by both the resonator and undamping. In order to adequately approach the quality of the resonator, the relative contribution of the undamping to the noise, i.e. the signal-dependent and signal-independent stochastic errors, has to be acceptably low. Further, the undamping must fulfill requirements with respect to the generation of harmonics of the reference signal. On the one hand, applications require *sinusoidal* signals and then harmonics are not allowed. On the other hand, an application may only be interested in the *repetition rate* of the signal; it does not matter whether it is a sine wave or a block wave, for instance. Thus, depending on the application, the specification for the signal-dependent systematic errors (weak and clipping distortion) may differ.

5.4 Signal power

Two types of LC-tank resonators can be distinguished:

- series resonators;
- parallel resonators.

The series resonance is characterized by a relatively low impedance whereas the impedance is relatively high for the parallel resonance. For both a series

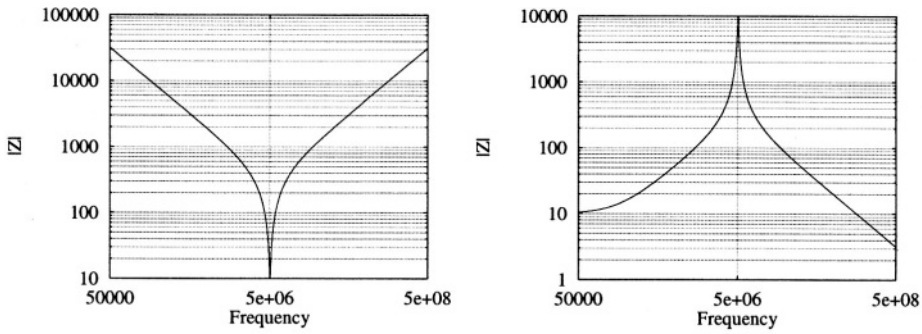

Figure 5.5: Typical plots of the impedance versus frequency of a series and a parallel resonator, $C_s = 100$ pF, $L_s = 10$ μH and $R_s = 10$ Ω.

and a parallel resonator, a typical plot of their impedance as a function of the frequency is depicted in figure 5.5. For the impedance at resonance it can easily be shown that the following holds:

$$Z(\omega_0)_{series} = R_s, \tag{5.15}$$

$$Z(\omega_0)_{parallel} = Q_s^2 R_s. \tag{5.16}$$

For the resonators depicted in figure 5.5, the quality factor equals about 32 and thus their impedances at resonance equal 10 Ω and $\approx$10 kΩ, respectively. For the series resonator, the power density of the equivalent input noise voltage, $S_{v-series}$, is given by:

$$S_{v-series} = 4kTR_s, \tag{5.17}$$

where T is the absolute temperature and k is the Boltzmann constant. For a given output signal, i_s, the carrier-to-noise ratio, CNR [dBc/Hz], yields [4]:

$$CNR(\Delta\omega) = 10\log\left[\frac{\hat{i}_s^2 R_s Q_s^2}{2kT} \cdot \left(\frac{\Delta\omega}{\omega_0}\right)^2\right], \tag{5.18}$$

where $\hat{i}$ is the amplitude of the sine wave, $\Delta\omega = \omega - \omega_0$. Thus the higher the quality factor, the higher the CNR.

Calculating the power-density spectrum of the equivalent input noise current, $S_{i-parallel}$, of the parallel resonator, for frequencies relatively close to ω_0, yields:

$$S_{i-parallel} = \frac{4kT}{Q_s^2 R_s}. \tag{5.19}$$

Calculating the obtainable CNR [dBc/Hz] for a given output signal yields the same expression as was found for the series resonator, equation (5.18). Thus,

with respect to power consumption and obtainable CNR ratio, it does not matter whether a series resonator or a parallel resonator is used. This conclusion is similar to the conclusion stated in Chapter 3 with respect to whether voltage or current domain processing has to be used in order to obtain the highest SNR for a given power consumption.

The difference between the series and parallel resonator is found when the voltage swing across their terminals is calculated. The ratio of the voltage swing for the parallel resonator, v_p, and the voltage swing for the series resonator, v_s, under the condition of equal power consumption and thus equal CNR, equals:

$$\frac{v_p}{v_s} = Q_s. \tag{5.20}$$

Thus, using a series resonator results in the lowest minimum required supply voltage. Therefore, in the remaining of this chapter the series resonator is used in the discussions. Of course, the calculations performed and conclusions found for the series resonator also apply for the parallel resonator.

5.5 Noise

In oscillator design a distinction is made in the electronic stochastic noise present in the circuits according to:

- amplitude noise;
- phase noise.

Or expressing it in the form of an equation for a reference signal $e(t)$:

$$e(t) = [A_0 + a_n(t)] * \sin[\omega_0 t + \phi_n(t) + \phi_0], \tag{5.21}$$

where A_0 is the mean amplitude, $a_n(t)$ is the amplitude noise, ω_0 is the frequency of the reference signal and $\phi_n(t)$ is the phase noise. The phase noise represents all the noise in the oscillator which influences the frequency of the reference signal. To obtain a high stability, this phase noise must be low. By definition, the amplitude noise does not influence the frequency. This does not mean that this amplitude noise may be arbitrarily large, because in subsequent blocks the amplitude noise can be converted into phase noise.

A typical plot of the output spectrum [5] of an oscillator is shown in figure 5.6. In the plot the power density of the oscillator output is drawn against the deviation from the carrier frequency. For frequencies relatively far from the carrier frequency, a white noise floor is found which is mostly the result of additive noise at the output of the resonator and the total oscillator. Closer to the carrier frequency, the phase noise increases by 20 dB/dec. This noise is mainly a result of white noise that is filtered by the resonator, i.e. due to white

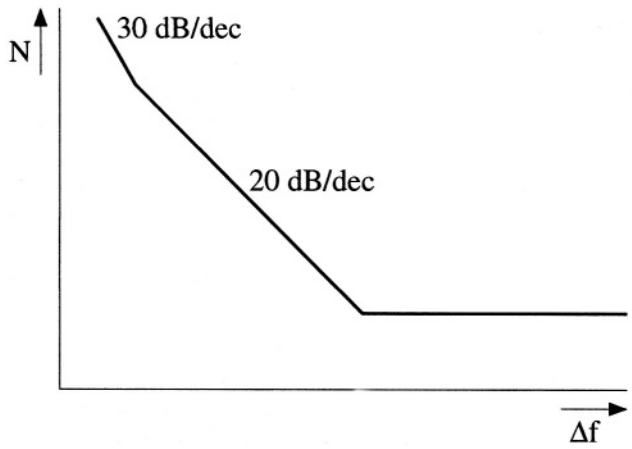

Figure 5.6: A log-log plot of a typical output spectrum of an oscillator.

noise at the input of the resonator, see equation (5.18). Going even closer to the carrier frequency, the noise increases by 30 dB/dec as a result of 1/f noise that is filtered by the resonator.

A common way to qualify an oscillator is by means of the carrier-to-noise-ratio. This is the ratio of the carrier power and the noise power density (phase noise *and* amplitude noise) found at a specified frequency (one sideband) from the carrier. As this frequency deviation is relatively smaller for higher oscillation frequencies than it is for a low-frequency oscillator, it cannot be used for comparing different oscillators. For this purpose the spectral density of fractional frequency fluctuations, the oscillator constant (OC) for short, is more convenient. It is defined according to:

$$\begin{aligned} OC &= 10\log\left[\mathcal{L}(\Delta\omega)\left(\frac{\Delta\omega}{\omega_0}\right)^2\right] \\ &= -CNR(\Delta\omega) - 3\text{ dB} + 20\log\left(\frac{\Delta\omega}{\omega_0}\right) \\ &= -CNR(\Delta\omega = \omega_0) - 3\text{ dB} \end{aligned} \tag{5.22}$$

where $\mathcal{L}(\Delta\omega)$ is the single-sideband phase-noise-to-carrier ratio. This measure is a constant as by multiplying $\mathcal{L}(\Delta\omega)$ by $(\Delta\omega)^2$ the phase-noise spectrum of -20 dB/dec is transformed into a flat spectrum. Further, by normalizing to the carrier frequency, relative frequency deviations are used when the power density of the phase noise is calculated. The oscillator constant can be seen as the extrapolated CNR at a distance of ω_0 from the carrier. The difference of 3 dB is due to the fact that the CNR also accounts for the amplitude noise, while $\mathcal{L}$ accounts for the phase noise only.

For oscillators using an LC-tank this constant can be as low as -180 dBc/Hz, whereas for high-performance crystal oscillators -220 dBc/Hz and even lower can be obtained[3].

Returning now to the resonator with its undamping, it was concluded that the performance of the undamping has to be such that it, ideally, does not degrade the performance of the intrinsic resonator. Of course, this can be obtained with a high power consumption in the undamping. With structured design techniques the degradation can be kept to a minimum without incurring unnecessary high power consumption, i.e. the power consumed is used efficiently.

In the rest of this chapter, the term noise is understood to mean the phase noise plus the amplitude noise of an oscillator. For additive noise superimposed on a sinusoidal signal, half of the noise power can be said to be phase noise while the other half is amplitude noise [4]. Calculations can therefore be performed on the total power of the noise sources; to find the corresponding phase noise one has to the divide the outcome by two.

An undamping can be realized in several ways. The most simple oscillator uses an active part supplying the required power for compensating for the losses in the resonator. The amplitude of the oscillation is bounded by non-linearities in this active part. The method as presented by Boon [4] uses an explicitly designed negative impedance which is realized with a double-loop feedback amplifier. The limiting is done explicitly by components in the feedback network. In this way the active part remains in its normal mode and no delays, et cetera, are introduced due to saturation effects. This method orthogonalizes the undamping and limiting action. As it fits conveniently in the structured design method as described in this book, it is used here for the discussions.

Consider the oscillator as depicted in figure 5.7. The undamping is realized by means of a negative impedance. This negative impedance is realized by means of the nullor with a double-loop feedback. The output of the nullor is used as the output of the oscillator, this is the additional port of the undamping already mentioned. For this oscillator a voltage output is chosen; however, a current output can be realized equally well. The gyrator feeds a current to the input as a function of the output voltage, whereas the transformer feeds a voltage to the input as a function of the same output voltage. Therefore a relation between the input voltage and input current is obtained, i.e. an accurate input impedance, Z_{in}, can be realized. Straightforward calculations yield:

$$Z_{in} = -\frac{1}{G \cdot n}, \tag{5.23}$$

where n is the transformer ratio and G is the transconductance of the gyrator. A negative impedance is found as one of the two feedback loops is a positive-

[3] The new cord-less telephone standard, CT-1, requires a *tunable* oscillator with an oscillator constant of -212 dBc/Hz. To realize this, orthogonalization has proven to be inevitable [2].

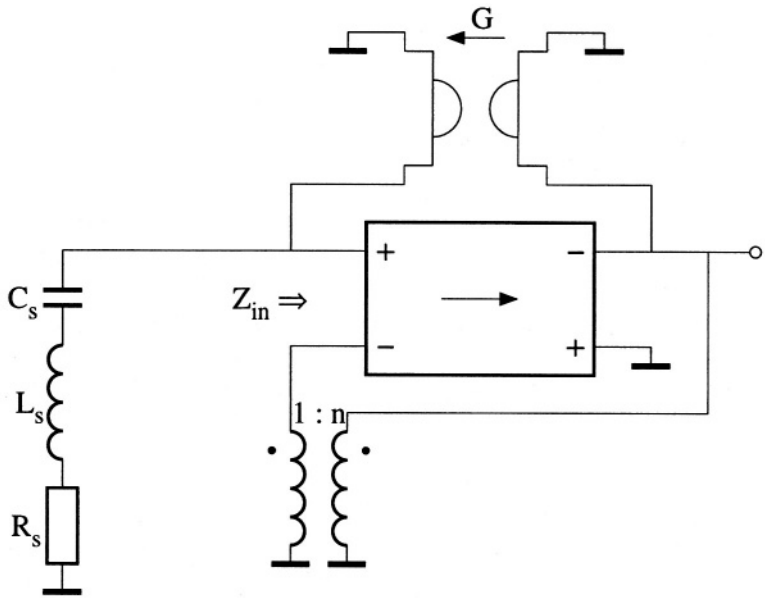

Figure 5.7: A series resonator undamped by means of a negative impedance realized by means of a nullor with a double-loop feedback.

feedback loop. According to network theoretical definitions, both feedback elements are ideal, i.e. noise free, et cetera, and thus the input impedance is also ideal. It must be noted however, that according to thermodynamics a gyrator cannot be noise free as it has a resistive transfer. Assuming the gyrator for the moment to be ideal, the influence of the noise of the input stage of the nullor implementation is studied.

In figure 5.8 the oscillator is depicted with the input stage noise sources, S_v and S_i. The input noise of the nullor implementation can be made negligible if S_v and S_i can be decreased enough. When a MOSFET is used as the input stage, the trivial solution is found that the drain current must be as large as possible. This is because the equivalent noise current and noise voltage both reduce for increasing drain current. For a bipolar input stage, the solution is less trivial and will be discussed here. To find the contribution of the bipolar transistor relative to the intrinsic noise of the resonator, the noise of the bipolar input stage is transformed into a single equivalent noise voltage source, which is in series with the noise voltage source of the resonator, S_{vres}. Straightforward calculations, assuming that the resonator is undamped by a negative resistance with a magnitude equal to R_s, yield:

$$S_{veq}(\omega) = S_{vres}(\omega) + S_{vbip}(\omega) + R_s^2 S_{ibip}(\omega), \tag{5.24}$$

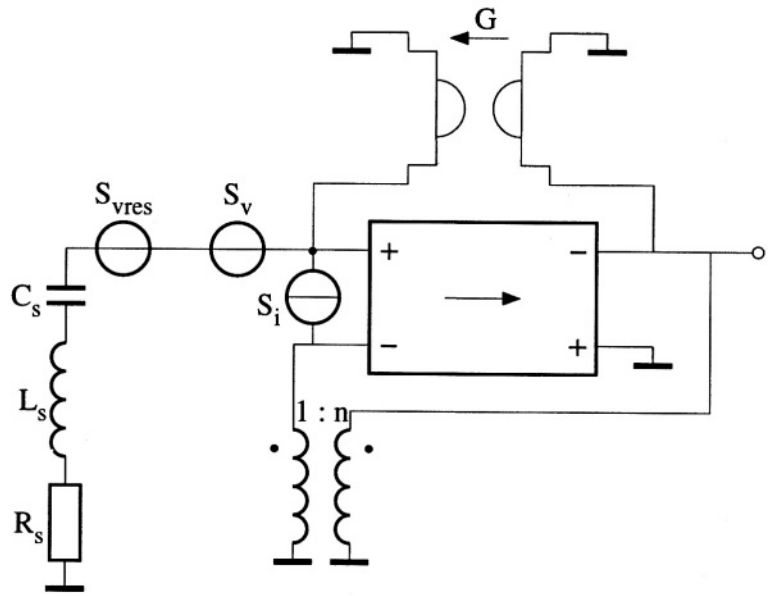

Figure 5.8: The equivalent noise sources, S_v and S_i, of the nullor implementation.

where S_{veq} is the power-density spectrum of the equivalent noise voltage source and S_{vbip}, S_{ibip} are the power-density spectra of the equivalent voltage noise and current noise sources of the bipolar transistor, respectively. In this expression, the term responsible for the noise floor is ignored.

It should be noted that the feedback networks do not have any influence on the noise performance, which is a result of the non-energetic gyrator and transformer [6]. In Section 5.7 an example is given of the design of an oscillator with realistic resistive feedback networks. Substituting the equivalent noise sources for the bipolar transistor, yields:

$$S_{veq} = 4kTR_s + 4kT\left(r_b + \frac{1}{2}\frac{1}{g_m}\right) + 2qI_BR_s^2, \tag{5.25}$$

where r_b is the base resistance of the transistor, g_m is its transconductance and I_B is its base current. In this equation the 1/f noise of the transistor is ignored for the sake of clarity. It can, however, be simply incorporated. Further, in the equation it is assumed that the frequencies of interest are relatively low, such that the current-gain factor of the transistor, β_f, can be assumed to be constant. To realize a relatively low transistor noise contribution of the transistor, r_b has to be smaller than R_s. As R_s can easily be on the order of several ohms to some tens of ohms, a relatively large transistor with a lot of base contacts has to

suffice. Further, $1/g_m$ has to be smaller than R_s, requiring a high bias current. Performing a minimization of the equivalent input noise level, with the collector current as the independent parameter, yields:

$$I_{Copt} = \frac{V_T\sqrt{\beta_f}}{R_s}. \tag{5.26}$$

Substituting this for the collector current in equation (5.25), yields for the equivalent input noise voltage:

$$S_{veq} = 4kT\left(R_s + r_b + \frac{R_s}{\sqrt{\beta_f}}\right). \tag{5.27}$$

Thus, the noise contribution of the bipolar transistor can be made negligible with respect to the resonator noise at the cost of a lot of power. For instance, when $V_T = 25$ mV, $\beta_f = 100$ and $R_s = 10\ \Omega$, the optimum collector current is 25 mA. But, as the base resistance is still directly in the expression, the base resistance is probably the dominant oscillator noise source.

The key problem is the low impedance of the series resonator. Using a transformer in order to increase this impedance level for a series resonator is a common technique [7]. However, transformers are very often too bulky. Tapping the resonator, either capacitively or inductively, is a powerful method for obtaining impedance transformations. Tapping the resonator can be used to obtain [4]:

- a power match;
- a noise match.

See for instance [8] in which an oscillator is described which uses a parallel resonator; a transformer is used to obtain a noise match and an inductive tap is used to obtain a power match such that the inherently large voltage swing of the parallel resonator is no longer directly limited by the power-supply voltage and breakdown voltages.

In [9] a series connection of several series resonators is proposed in order to obtain a *power* match. By using the voltage across the capacitance of one resonator, higher internal voltages can be obtained such that they are no longer directly limited by the supply voltages or breakdown voltages.

In the following sections, capacitive tapping of the series resonator is shown to be a powerful method for obtaining a noise match to the undamping. Further, the limits to this tapping are described.

5.5.1 Tapping

The principle of a capacitively tapped resonator is depicted in figure 5.9a and its equivalent circuit diagram is depicted in figure 5.9b [10]. The capacitive

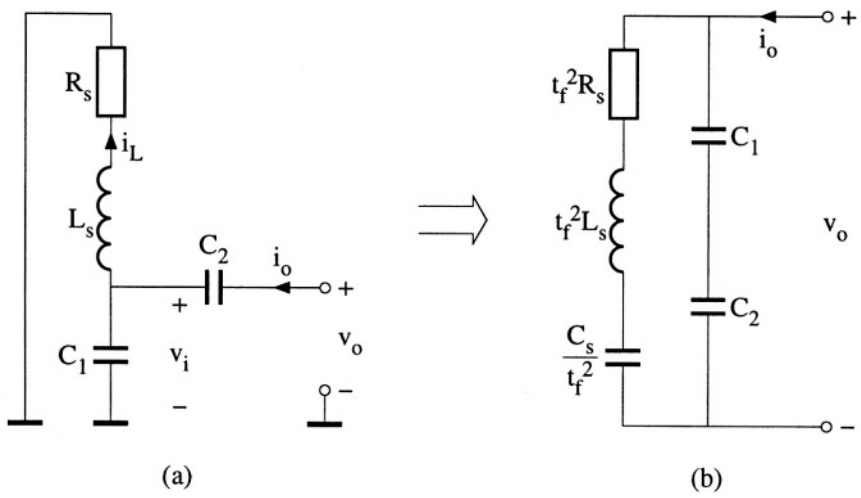

Figure 5.9: a) A capacitively tapped resonator and b) its equivalent circuit diagram, $C_s = C_1 + C_2$.

tap results in an impedance increase which can be understood as follows. The starting point is the notion that the inductor current for the tapped resonator is equal to the inductor current of the non-tapped resonator. Then the impedance can be calculated from the output current of the resonator and the *required* input voltage of the resonator. The ratio of the inductor current, i_L, and the resonator output current, i_o, is given by (assuming a relatively low undamping impedance compared with the impedance of C_2):

$$\frac{i_o}{i_L} = \frac{C_2}{C_1 + C_2}, \tag{5.28}$$

whereas the relation between the inductor current and the *required* input voltage is given by:

$$\frac{i_L}{v_i} = \frac{C_1 + C_2}{C_2} \cdot \frac{1}{Z_s}, \tag{5.29}$$

where Z_s is the impedance of the non-tapped resonator. Eliminating the inductor current from these two expressions yields:

$$Z_s' = \frac{v_i}{i_o} = \left(1 + \frac{C_1}{C_2}\right)^2 Z_s = t_f^2 Z_s, \tag{5.30}$$

where Z_s' is the impedance of the tapped resonator and the tap factor, t_f, is defined as:

$$t_f = 1 + \frac{C_1}{C_2}. \tag{5.31}$$

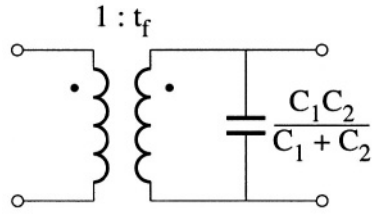

Figure 5.10: The circuit equivalent of capacitive tapping.

Thus the impedance level increases by t_f^2. This is also found when the equivalent circuit diagram is derived [10]. The equivalent resonator consists of components that are scaled such that their impedance increases by a factor t_f^2. The resonance frequency is not changed, which also holds for the quality factor of the inductor. Parallel to the equivalent series resonator, the series connection of the two capacitors used to make the tap is found [4]. For the tapped resonator, the impedance for relatively high frequencies is determined by this series connection of C_1 and C_2, which follows straightforwardly from both figures. Thus capacitive tapping is almost identical to an ideal transformer, except for the additional parallel capacitance, see figure 5.10. Due to this additional capacitance, the impedance is not exactly real at $\omega_0 = 1/\sqrt{L_s C_s}$. With straightforward calculations the impedance of the *tapped* resonator, Z_s', at ω_0 can be found to be equal to:

$$Z_s'(\omega_0) = \frac{t_f^2 R_s}{1 + \left(\frac{t_f - 1}{Q}\right)}. \tag{5.32}$$

Which can be approximated by $t_f^2 R_s$ when the tap factor is much smaller than the quality factor of the resonator.

For the noise behavior it is assumed that R_s is the only source generating noise, i.e. $S_v = 4kTR_s$, see figure 5.11a. This noise source has to be transformed to the output of the resonator, the current, in order to compare it to the non-tapped resonator. Performing this transformation on the principle circuit diagram, figure 5.11a, yields the following for the power density of the equivalent noise current, S_{ieq}:

$$S_{ieq} = \frac{4kTR_s}{|Z_s|^2 t_f^2}, \tag{5.33}$$

where Z_s is the impedance of the *non-tapped* resonator. Thus, the tapping can be seen as an ideal transformation for the noise. That this is true can be elucidated with the help of figure 5.11b in which the noise is assumed to be

[4] These capacitors were not found from the previous simple calculations due to the assumption which was made before equation (5.28).

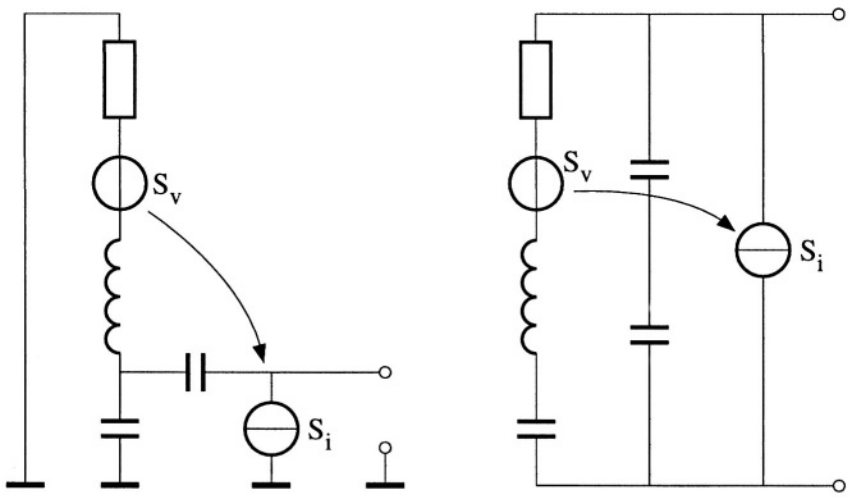

Figure 5.11: Transformation of the noise of the tapped resonator to an equivalent noise current. a) for the actual circuit diagram and b) for the equivalent circuit diagram.

generated by the resistor $t_f^2 R_s$, as this is the only resistive part of the equivalent circuit. Transforming this noise voltage to a noise current by means of the Thévenin-Norton transformation with the equivalent series resonator as the source impedance, i.e. the scaled impedance, yields:

$$S_{ieq} = \frac{4kTt_f^2 R_s}{\left(|Z_s| \cdot t_f^2\right)^2}, \tag{5.34}$$

which is the same as equation (5.33).

Thus the power-density spectrum of the equivalent noise current is equal to the original one, except for the scaling of $1/t_f^2$. But, as the output current of the resonator is decreased by t_f the CNR for a given resonator power equals:

$$\begin{aligned} CNR(\Delta\omega) &= 10\log\left[\frac{\frac{1}{2}\hat{i}_{out}^2}{\frac{kTR_s}{Q_s^2 R_s^2 t_f^2}\left(\frac{\omega}{\Delta\omega}\right)^2}\right] \\ &= 10\log\left[\frac{\hat{i}_{out}^2/t_f^2 R_s Q_s^2}{\frac{2kT}{t_f^2}}\left(\frac{\Delta\omega}{\omega}\right)^2\right] \end{aligned} \tag{5.35}$$

$$= 10 \log \left[\frac{\hat{i}_{out}^2 R_s Q_s^2}{2kT} \left(\frac{\Delta\omega}{\omega} \right)^2 \right],$$

where i'_{out} and i_{out} are the output current of the tapped and non-tapped resonator, respectively. The maximum attainable CNR does not change due to the tapping, see equation (5.18).

5.5.2 CNR maximization

In the expression for the CNR ratio, equation (5.35), not only the noise is of importance, but also the stubbornness of the resonator for that noise, indicated by the quality factor. Reducing the influence of the noise of the active part can be done by tapping, see Section 5.5.3. For higher tap factors the impedance level increases and the relative contribution of the active part to the total noise power reduces. But, the higher the tap factor is, the weaker the coupling of the resonator to the undamping is via capacitor C_2, which becomes relatively small for relatively high tap factors. Therefore it may be expected that the resonator loses grip on the oscillator; it can no longer force its frequency selectivity onto the oscillator and thus the quality factor reduces. This reduction of the quality factor may cancel the CNR improvement which was obtained by minimizing the noise level or even surpass it, such that the CNR reduces.

The most straightforward way is to write down the complete expression, with all the dependencies of the noise and the quality factor on the tap factor, and optimize it. However, this is a multi-dimensional optimization not giving much insight into the real problem.

Therefore, orthogonality is assumed between those two aspects and they are separately treated. After that, a check can be made as to the extent that this orthogonality was permissible.

5.5.3 Noise minimization

With tapping it is possible to do an impedance transformation without degrading the intrinsic CNR ratio for a given power consumption of the resonator. Incorporating the tapped resonator in the oscillator of the figure 5.8, yields the schematic of figure 5.12. To be able to show the effect of tapping on the CNR in a more straightforward way, the equivalent output noise current is used instead of the equivalent input noise voltage. Further, it is assumed that the resonator is undamped by an impedance equal to $t_f^2 R_s$. Then, the expression is given by, ignoring again the term for the noise floor:

$$S'_{ieq} = \frac{4kT}{t_f^2 R_s} + 2qI_B + \frac{4kT\left(\frac{1}{2}\frac{1}{g_m} + r_b\right)}{\left(t_f^2 R_s\right)^2}. \tag{5.36}$$

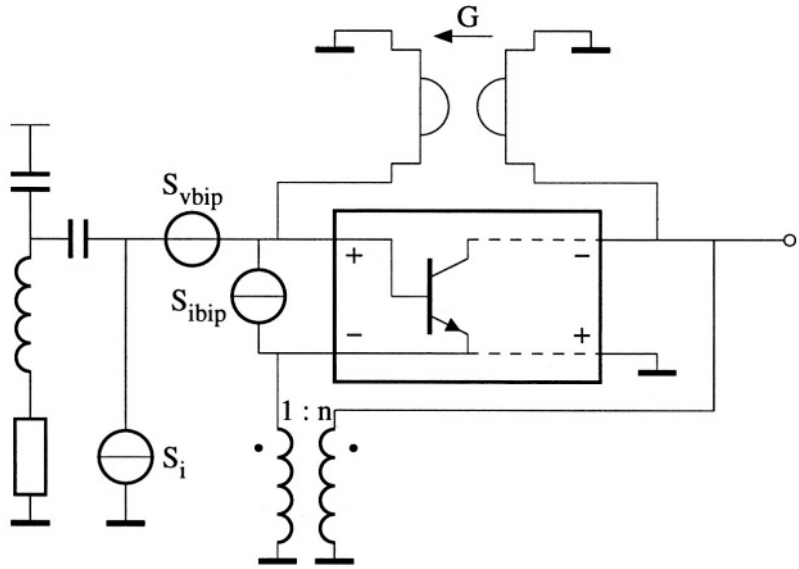

Figure 5.12: The oscillator using a tapped resonator with a bipolar transistor as the input of the nullor implementation.

From this expression, the optimum collector current can also be found, for which the noise contribution of the transistor is minimum. The optimum collector current is, again, given by equation (5.26) with R_s replaced by $t_f^2 R_s$. The noise-power density of the equivalent noise current, for the optimum collector current, is given by:

$$S_{ieq} = 4kT \left[\frac{1}{t_f^2 R_s} + \frac{r_b}{\left(t_f^2 R_s\right)^2} + \frac{1}{t_f^2 R_s \sqrt{\beta}} \right]. \tag{5.37}$$

This expression closely resembles equation (5.27). Besides that, it is expressed in the current domain instead of the voltage domain; with the tap factor the influence of the base resistance can now be reduced. The relative influence of the noise due to the base and collector shot noise remain the same. This expression is depicted in figure 5.13, including the signal output power (ignore the dotted line for the moment). For an increasing tap factor, the power of the output signal reduces quadratically, which also holds for the resonator thermal noise and the transistor base and collector shot noise. In contrast, the noise contribution of the base resistance reduces with the fourth power of the tap factor. Therefore, beyond a certain tap factor, the influence of the base resistance can be made negligibly small. Thus, with tapping the constraint of a relatively low base resistance can be relieved to a large extent. For a current-gain factor of about a

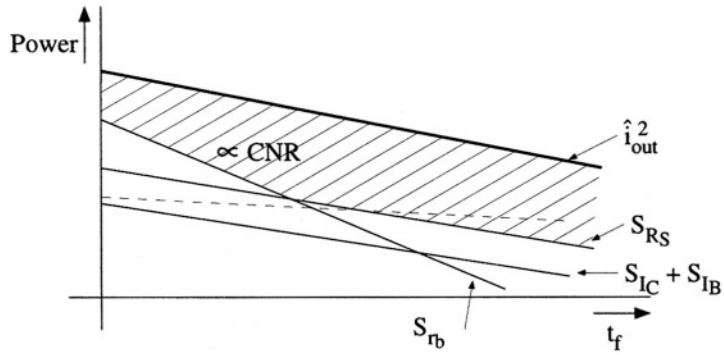

Figure 5.13: The CNR of an oscillator as a function of the tap factor.

hundred, the CNR reduces by only 0.4 dB, see equation (5.37). Increasing the tap factor even further, the CNR no longer increase but the impedance level still increases and the optimum collector current reduces to zero, see equation (5.26) again with R_s replaced by $t_f^2 R_s$, thus saving power.

The tap factor should be infinite with respect to the noise and power consumption. This is mainly a result of the ideal feedback networks. In Section 5.7 a practical example is discussed in which noise is introduced by a feedback element, of which the noise power is only *proportional* to the tap factor, indicated by the dotted line in figure 5.13. This results in a tap factor, not being infinite, for which the CNR is maximal.

5.5.4 Q degradation

In this section the dependency of the quality factor on the tap factor is discussed. Examining the equivalent diagram of the tapped resonator (figure 5.9), shows that the resonator is inherently deteriorated by a parallel resonance. The impedance level of the equivalent resonator increases proportionally to the square of the tap factor, whereas the impedance level of the two series connected capacitances, which are in parallel with the resonator, increases approximately proportionally to the tap factor only, i.e.:

$$\frac{C_1 C_2}{C_1 + C_2} = C_s \left(\frac{1}{t_f} - \frac{1}{t_f^2} \right) \approx \frac{C_s}{t_f}. \tag{5.38}$$

Therefore, it can be expected that the parallel resonance will get more influence at higher tap factors. This is illustrated in figure 5.14. The quality factor of the resonator is related to the phase-frequency plot according to:

$$Q_s = \frac{1}{2} \omega_{osc} \frac{d\phi}{d\omega}|_{\omega_{osc}}. \tag{5.39}$$

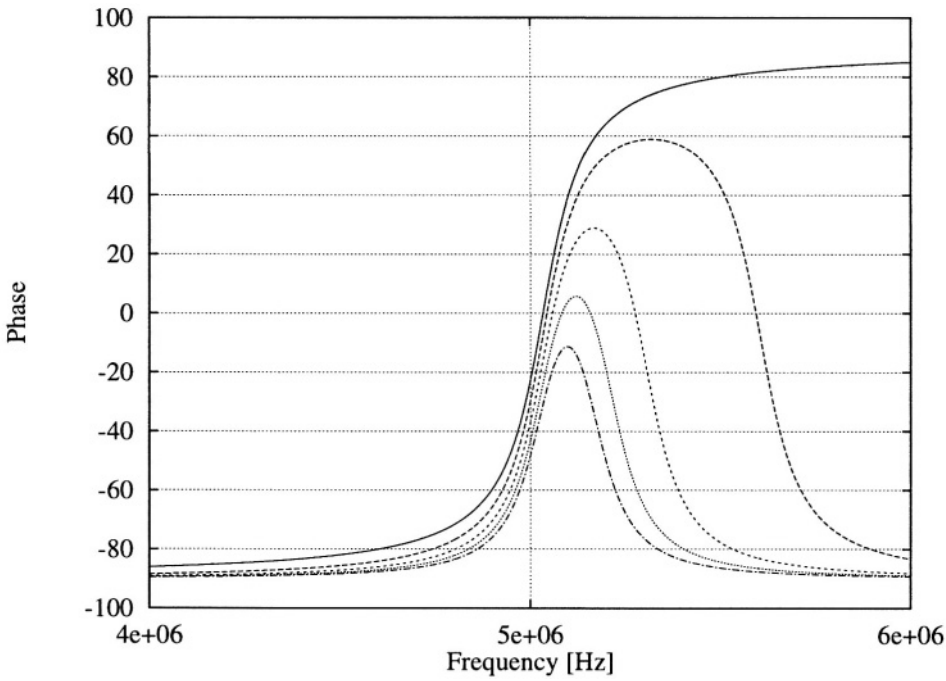

Figure 5.14: The phase shift as a function of the frequency with $t_f = 1$, $\frac{1}{6}Q_s$, $\frac{1}{3}Q_s$, $\frac{1}{2}Q_s$, $\frac{2}{3}Q_s$ corresponding to the functions ordered from the upper one to the lower one.

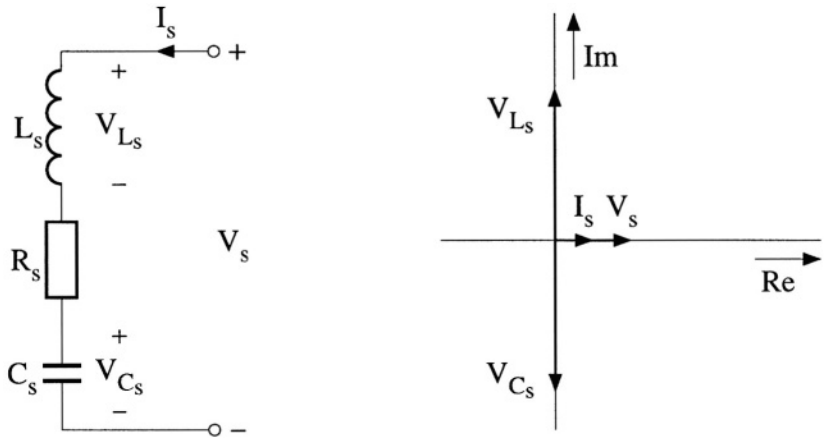

Figure 5.15: The non-tapped resonator and its vector diagram.

Thus the quality factor depends on the steepness of the frequency-versus-phase plot at ω_{osc}, the oscillation frequency. In the figure is clearly visible that for higher tap factors the parallel resonance shifts to the series resonance and the steep slope of the frequency-versus-phase function completely vanishes. Thus the quality factor reduces.

To find the relation between tapping and Q degradation, the original non-tapped resonator is again considered. In figure 5.15 the non-tapped resonator and its vector diagram are depicted. When oscillating, the impedance of the resonator is real and thus the input voltage and current are in phase. The capacitor and inductor current are, respectively, –90° and +90° shifted with respect to the output voltage. Further, due to the resonance the capacitor and inductor voltage are much larger than the resulting output voltage. For the tapped resonator the equivalent schematic and its vector diagram are depicted in figure 5.16. Again it is assumed that the resonator oscillates because it is undamped by means of a negative real impedance. The phase of the output voltage and current is used as the reference phase. The input current is a result of I_{cap} through the series connection of the two tap capacitors and the current through the equivalent resonator, I_{R_t}. In order to get an output current which is in phase with the output voltage (recall the real undamp impedance), the current I_{R_t} has to become slightly inductive in order to cancel the capacitive current, I_{cap}. Thus, the frequency of oscillation will increase slightly due to the tapping in the case of a real undamping. This frequency shift is required to calculate the phase detuning of the resonator and from that the Q degradation.

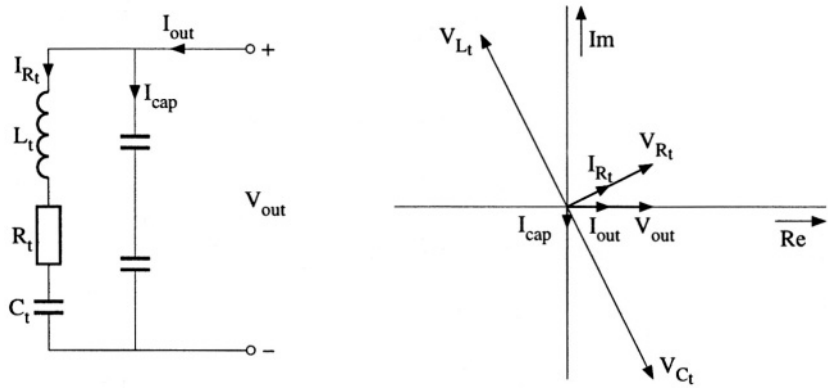

Figure 5.16: The tapped resonator and its vector diagram.

5.5.4.1 Frequency shift due to tap

The constraint for a real impedance of the tapped resonator in order to oscillate due to a negative real undamping impedance is:

$$\mathrm{Im}(I_{R_t}) = -I_{cap}, \tag{5.40}$$

where Im denotes the imaginary part of a current. The imaginary part of the equivalent resonator current equals:

$$\mathrm{Im}(I_{R_t}) = V_{out} \frac{\left(\frac{1}{\omega C_t} - \omega L_t\right)}{R_t^2 \left[\left(\frac{1}{\omega R_t C_t} - \frac{\omega L_t}{R_t}\right)^2 + 1\right]}. \tag{5.41}$$

When it holds that:

$$Q_s^2 \left(\frac{\omega_0}{\omega} - \frac{\omega}{\omega_0}\right)^2 \ll 1, \tag{5.42}$$

where $\omega_0 = \frac{1}{\sqrt{L_t C_t}}$, equation (5.41) can be approximated by:

$$\mathrm{Im}(I_{R_t}) = \frac{V_{out}}{R_t^2} \left(\frac{1}{\omega C_t} - \omega L_t\right). \tag{5.43}$$

The constraint of equation (5.42) seems to be less fulfilled for higher Q_ss. However, in a few moments it will be shown that the contrary holds, i.e. the constraint is *better* met for higher Q_ss. Solving equation (5.40) for the new oscillation frequency, ω_{osc}, yields:

$$\omega_{osc}^2 = \frac{1}{L_t C_t \left(1 - \frac{R_t^2 C_p}{L_t}\right)}, \tag{5.44}$$

where $C_p = \frac{C_1 C_2}{C_1 + C_2}$. Introducing the tap factor and the quality factor in this expression, results in:

$$\omega_{osc}^2 = \frac{1}{L_t C_t \left[1 - \frac{1}{Q_s^2}(t_f - 1)\right]}. \tag{5.45}$$

From this equation it follows that for higher tap factors the frequency of oscillation increases slightly. For a tap factor which is about 10% of the quality factor, this increase is only 0.05%. Now the frequency of oscillation is found, the constraint from equation (5.42) can be checked. Substitution of ω_{osc} for ω in equation (5.42) yields:

$$\frac{(t_f - 1)^2}{Q_s^2 - (t_f - 1)} \ll 1, \tag{5.46}$$

which holds when:

$$(t_f - 1) \ll Q_s. \tag{5.47}$$

When this criterion is met, the expression for the frequency of oscillation is valid.

This frequency shift is relatively small as a result of the fact that the steepness of the phase-versus-frequency plot of the resonator is much larger than the steepness of the phase-versus-frequency plot of the two series-connected capacitors.

5.5.4.2 The phase shift due to the tap

As a result of the last conclusion of the previous section, the phase shift of the two series-connected capacitors may be assumed to be independent of the frequency, in the range of frequencies which are relatively close to ω_0. Thus for calculating the additional phase shift that the resonator (L_t, C_t, R_t in figure 5.16) has to introduce in order to compensate for the phase shift due to the two series-connected capacitors, the frequency may be assumed to be ω_0. Subsequently, from the relation between the phase shift and the quality factor, the Q degradation can be obtained.

Thus, the frequency of oscillation is assumed to be:

$$\omega_{osc} = \frac{1}{\sqrt{L_t C_t}}. \tag{5.48}$$

In this case the inductor and capacitor impedance cancel each other out and the equivalent circuit of figure 5.17 can be used. The phase shift, ϕ, between the output voltage and current can be found by straightforward calculations, yielding:

$$\phi = \arctan\left(\frac{t_f - 1}{Q_s}\right). \tag{5.49}$$

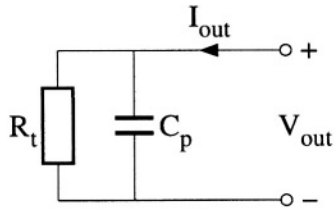

Figure 5.17: The equivalent circuit diagram of the tapped resonator at $\omega = \omega_{osc}$.

Of course, the influence of the tapping is again less for higher Q_ss and less tapping. For a tap factor which is about 10 % of the Q_s, the phase shift is 0.09 radians.

5.5.4.3 The Q degradation due to a phase shift

The quality factor of a resonator is a measure for indicating the sensitivity of the frequency of oscillation to phase variations. The higher the quality factor is, the less frequency fluctuations arise from phase fluctuations. For the maximum quality factor of a resonator, the simplified part of equation (5.7) can be used. However, to obtain an expression of the quality factor as a function of the phase shift, the principal expression has to be used:

$$Q_s = \frac{1}{2}\omega_{osc}\frac{d\phi}{d\omega}|_{\omega_{osc}}. \tag{5.50}$$

Calculating this for the non-tapped resonator, see Appendix C, yields:

$$Q_{eff} = \cos^2\phi \cdot Q_s \approx \left(1-\phi^2\right)^2 \cdot Q_s \tag{5.51}$$

where Q_{eff} is the effective quality factor and Q_s is the maximum quality factor of the resonator, i.e. with ϕ=0. This expression resembles the expression as given in [11]. However, the calculation done in Appendix C is believed to be more straightforward. For phase shifts $\phi \ll 1$, the expression simplifies to:

$$Q_{eff} \approx \left(1-\phi^2\right)^2 \cdot Q_s \tag{5.52}$$

5.5.4.4 Q degradation due to the tapping

Now that expressions have been found for the phase shift as a function of the tap factor, equation (5.49), and the effective quality factor as a function of the phase shift, equation (5.51), the effective quality factor as a function of the tap factor can be found. This expression is readily found to be:

$$Q_{eff} = Q_s \cdot \frac{1}{1+\left(\frac{t_f-1}{Q_s}\right)^2}. \tag{5.53}$$

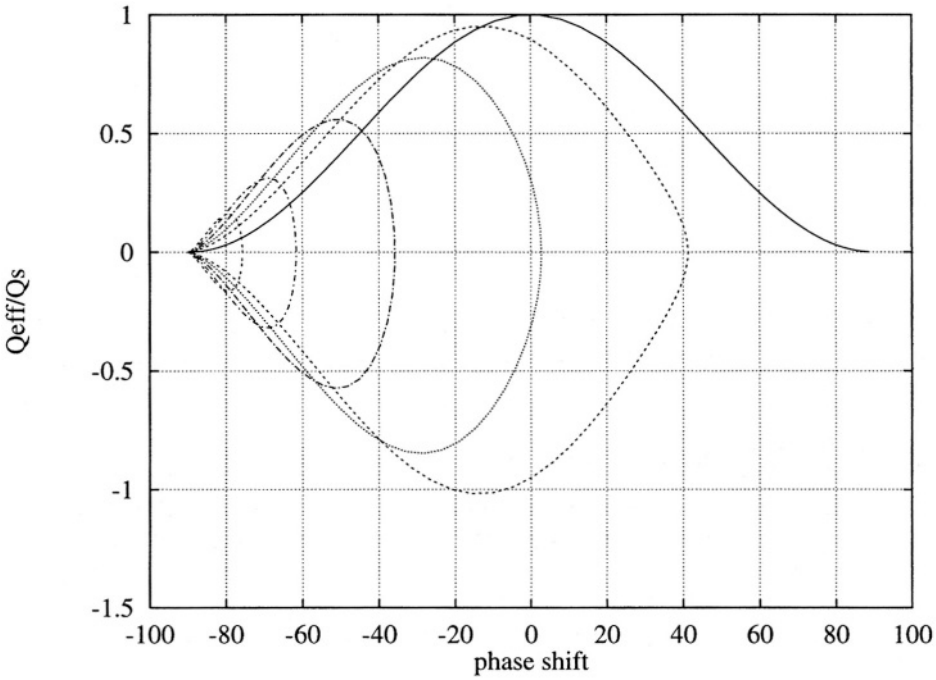

Figure 5.18: The relative change in the quality factor as a function of the phase shift with the tap factor as a parameter. For the functions the tap factor equals, from the right to the left, respectively, $t_f = 1$, $\frac{1}{4}Q_s$, $\frac{1}{2}Q_s$, Q_s, $2Q_s$ and $4Q_s$.

From this equation it can easily be seen that for tap factors which are low compared to the quality factor of the resonator, the effective quality factor is only slightly lower than the intrinsic quality factor of the resonator. For instance, for a tap factor which is 10 % of Q_s, the effective Q is $0.992 \cdot Q_s$. In figure 5.18, the relative change in quality factor is depicted as a function of the phase shift with the tap factor as a parameter. From the figure it can be seen that for t_f=1 the plot has the shape of a $\cos^2 \phi$ function, i.e. equation (5.51). Further, for t_f=1 the plot is only found at "positive" Qs as no parasitic parallel resonance is found. When tapping, a parasitic parallel resonance is found and the plots also show "negative" Qs. The Qs becomes "negative" for the parallel resonance as the definition for the series resonator is used. Consequently, a minus sign arises as the derivative of the frequency as a function of the phase at the parallel resonance is negative, see equation (5.50) and figure 5.14.

The intersection point of the several functions with the $\cos^2 \phi$ are given by equation (5.49). As can be seen, this intersection point for higher tap factors

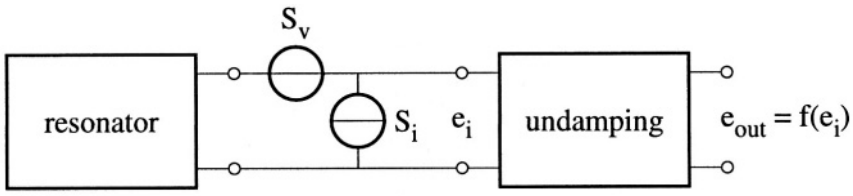

Figure 5.19: Block diagram of the oscillator for studying the noise floor

is no longer found at the maximum Q_{eff} for a given tap factor. However, this appears to be for relatively large tap factors.

Summarizing, it can be said that the Q-degradation due to tapping can be said to be negligible when the tap factor is much smaller than the quality factor. Thus when the tap factor resulting from noise minimization is much less than the quality factor, it was permissible to treat the Q degradation and the noise minimization orthogonally.

5.5.5 Influence on the noise floor

For relatively high and low frequencies, compared with the carrier frequency, the noise spectrum becomes flat again, i.e. the noise floor. The influence of tapping on this noise floor is studied in this section. Consider the block diagram of an oscillator as depicted in figure 5.19. The two noise sources depicted represent the total noise in the oscillator.

First, the situation for the non-tapped resonator is dealt with. The noise floor, as was said, is found at frequencies relatively high and low compared with the carrier frequency. In these frequency ranges, the impedance of the resonator is relatively high and the equivalent voltage noise does not affect the noise level at the output of the oscillator. In contrast, as a result of the relatively high resonator impedance, the equivalent noise current is completely transformed to the output of the oscillator and thus the CNR [dBc/Hz] at the noise floor is given by:

$$CNR = 10 \log \left(\frac{\hat{i}_{out}^2}{2S_i} \right), \tag{5.54}$$

where $\hat{i}_{out}$ is the amplitude of the resonator output current.

For the tapped resonator, the situation at relatively low frequencies is the same compared with the situation for the non-tapped resonator. Due to the relatively high resonator impedance, the equivalent noise current is also completely transformed to the output. But as the relative contribution of the noise of the active part is reduced by the tapping, see figure 5.13, a higher CNR for this noise floor may be found. For relatively high frequencies the situation is

different. As the resonator impedance becomes relatively low again, the equivalent noise current is shorted by the resonator impedance. Now the equivalent noise voltage determines the noise floor. But again, as the relative contribution of the noise of the active part is reduced by the tapping, a higher CNR may be found.

When non-ideal feedback networks are examined, the resulting noise floor increases as a result of the different dependencies of their noise level on the tap factor, see figure 5.13. But the noise level may still be below the noise floor for the oscillator with the non-tapped resonator. If not, then with relatively simple filtering compared with the filtering of the resonator itself, this noise floor can be reduced again, when it is unacceptably high.

5.6 Bandwidth

Applying the undamping schematic as depicted in figure 5.7, a high-frequency stability problem occurs. For relatively high frequencies, the tapped resonator can be considered as being only the series connection of C_1 and C_2, see figure 5.9b. As the bandwidth of the undamping impedance is still infinite (as a result of the nullor), a pole in the right-half plane is found at:

$$p = +\frac{1}{\frac{C_1 C_2}{C_1+C_2} R_t} = \omega_{osc} \frac{Q_s}{t_f - 1}. \tag{5.55}$$

where R_t is the magnitude of the undamping resistance $(R_t = t_f^2 R_s)$, see figure 5.16. Something analogous to this right-half plane pole is encountered in [12], in which a parallel parasitic capacitance resulted in parasitic oscillations. This is, of course, not permissible.

5.6.1 Compensation of the parallel C

To prevent the oscillator from parasitic oscillation, the tapped resonator should not only be undamped by means of a real impedance, but a negative capacitive part also has to be included. This is depicted in figure 5.20. The parallel C due to the tapping is compensated for by a negative capacitance. The instability problem is now solved.

Now that the parallel capacitance has been compensated for, it should be noted that the quality factor of the resonator is Q_s again when the frequency-versus-phase plot is used to calculated the quality factor. Thus for disturbances from outside the oscillator it once more reacts as an oscillator with a quality factor of Q_s. This does not hold for all the disturbances from inside the oscillator. For the noise due to the negative R, a resonator Q of Q_s is seen as this noise sources encounters a compensated parallel C. The noise of the actively

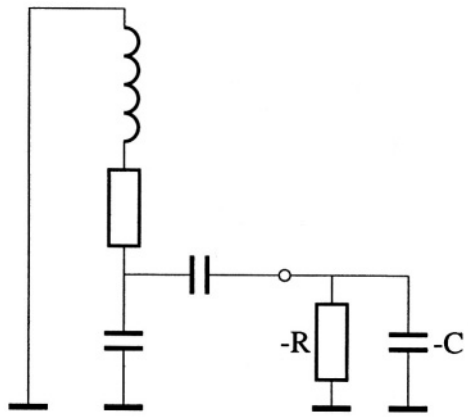

Figure 5.20: Undamping the tapped resonator with both a negative R and a negative C

realized negative capacitance, however, does not experience an impedance of an intrinsic resonator.

Similar considerations hold for the power consumption. For the tapped resonator without C compensation, the reactive power in the parallel capacitance resulted in an increase in the power consumption, the power of the parallel capacitance was dissipated by the resonator and the undamping resistance. With the compensating capacitance, the reactive power is no longer supplied and dissipated by the resonator, but the negative C supplies and dissipates the power for the parallel C. As this (active) negative C is not as efficient as the resonator, the power consumption increases.

Figure 5.21 shows an implementation of an undamping impedance in which the negative R and C are combined. The input impedance is readily obtained from the schematic and equals:

$$Z_{undamping} = -\frac{R_1 R_2}{R_3} // -\frac{R_3}{R_2} C_1. \tag{5.56}$$

5.6.2 Bandwidth of the undamping impedance

When the nullor of the undamping impedance is implemented, the bandwidth of the impedance is no longer infinite and additional phase shifts are obtained. This is depicted in figure 5.22. For the effective quality factor of the oscillator as a function of the bandwidth and oscillator frequency, the following expression holds:

$$Q_{eff} = \frac{Q_s}{1 + \left(\frac{f_{osc}}{B}\right)^2}, \tag{5.57}$$

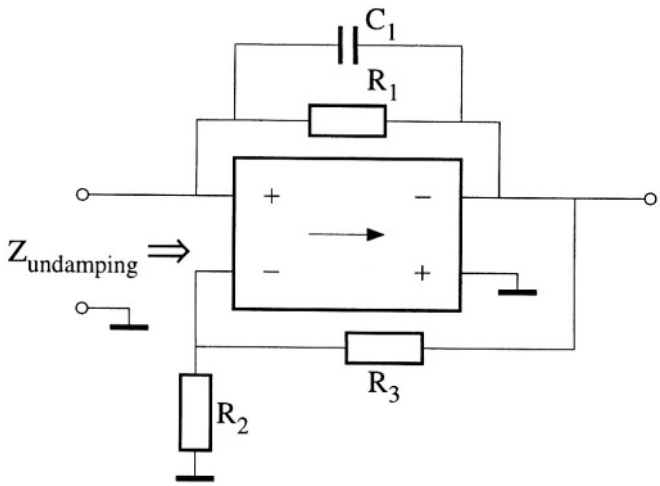

Figure 5.21: An implementation of the negative R and C, combined in one double-loop amplifier.

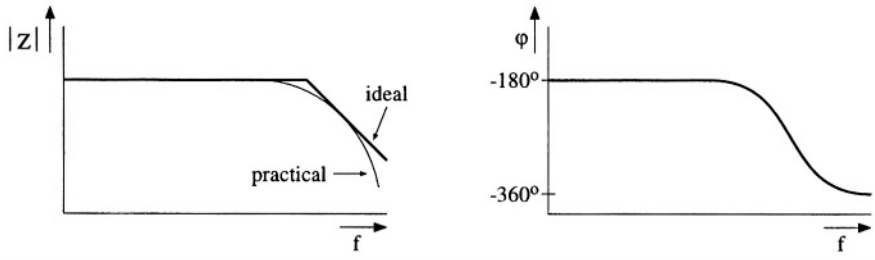

Figure 5.22: The bandwidth of the undamping impedance.

where B is the bandwidth of the undamp impedance with a first-order frequency behavior. This expression can be found by substituting the phase shift resulting from the limit bandwidth into the expression for the effective quality factor, equation (5.51), and rewriting the expression obtained.

For instance, when the bandwidth is double the oscillation frequency, 20 % of the resonator Q is lost.

Two design philosophies at circuit level can be used to minimize this effect to an acceptable level:

- maximize bandwidth;
- use the finite bandwidth to realize the negative C.

Maximize bandwidth In this case the model as described in [4] can be used. It is based on the dependence of the input impedance of an amplifier on the loop gain as described in [13] and adapted to the asymptotic-gain model in [6]. The input impedance is described as:

$$Z_{in} = Z_{in_\infty} \frac{-A\beta_{op}}{1 - A\beta op} \cdot \frac{1 - A\beta_{sc}}{A\beta_{sc}} \tag{5.58}$$

where Z_{in_∞} is the input impedance when the active part is a nullor, $A\beta_{op}$ is the loop gain of the amplifier when the input port is open and $A\beta_{sc}$ is the loop gain of the amplifier when the input port is short circuited. The frequency behavior of $1 - A\beta_{op}$ and $1 - A\beta_{sc}$ can be designed by means of the methods described in Chapter 4.

Use the finite bandwidth to realize the negative C As can be seen from figure 5.22 the undamping impedance with an ideal bandwidth also has a phase shift at the resonance frequency. This phase shift is a result of a reactive element in the feedback network in order to realize a negative C, see for instance figure 5.21. This equivalent negative C is also obtained when the bandwidth of the amplifier is limited. Thus, instead of realizing an amplifier with an accurate input impedance ($-R \,//\, -C$) over a relatively large bandwidth, it may be more easy to realize a negative R with a limited bandwidth.

5.7 Design examples

To show the positive influence of tapping on the CNR of an oscillator, the design of, and the measurements on two oscillators are discussed. One oscillator uses a non-tapped resonator whereas the other uses a tapped resonator. Before discussing the two designs, common specifications and design steps are discussed first.

C_s	560 pF
L_s	47 μH
R_s	6 Ω
f_{res}	981 kHz
$Q_s@f_{res}$	48

Table 5.1: Characteristic values of the chosen resonator.

5.7.1 Common topics and specifications

The core of both oscillators is a:

- Series resonator.

The resonator is realized by means of a discrete inductor and a capacitor. The inductors used have a maximum quality factor on the order of 50. This quality factor is attained at frequencies on the order of one to a few MHz. As the tap factor of a resonator has to be considerably smaller than its quality factor, see for instance equation (5.53), the frequency of oscillation is chosen to be:

- $f_{osc} = 1$ MHz,

so that this maximum quality factor is obtained. The values for the inductor and capacitor, which are chosen to be conform the E-12 series, and some other parameters are listed in table 5.1. The minimum attainable oscillator constant, *OC,* for this resonator is given by, see equation (5.18) and (5.22):

$$OC = -10 \log \left(\frac{I_s^2 R_s Q_s^2}{2kT} \right) - 3 \text{ dB}, \tag{5.59}$$

where I_s is the amplitude of the resonator current. In this expression it is assumed that half of the noise power is converted into phase noise. In order to ease the measurements to be done, the resonator current is chosen to be about:

- $I_s = 10$ μA,

resulting in a minimum oscillator constant of:

- $OC_{min} = -146$ dBc/Hz.

The configuration of the active part of both oscillators is as depicted in figure 5.23. For the output signal a voltage is chosen in order to facilitate measurements. To account for some loading of measurement equipment, a buffer amplifier or a probe, for instance, the following values for the load are chosen:

- $Z_{load} = R_{load} // C_{load} = 1$ MΩ//10 pF.

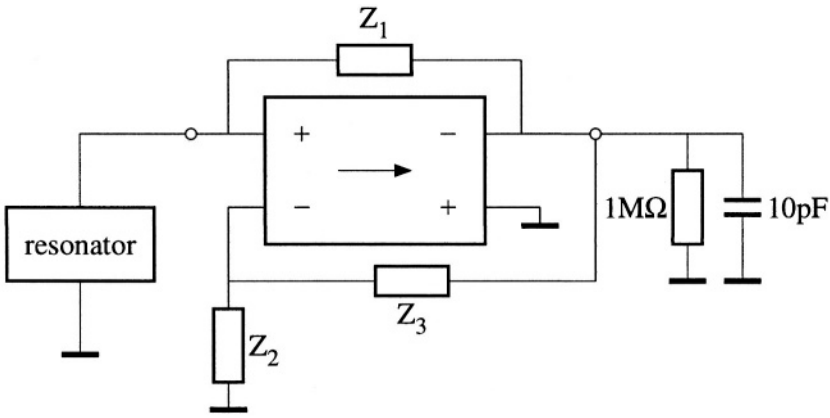

Figure 5.23: The basic configuration of the active part.

The input impedance, Z_{in}, of this structure is given by:

$$Z_{in} = -Z_1 \frac{Z_2}{Z_3}. \tag{5.60}$$

The values of these impedances are chosen on the basis of the noise and the signal power. The output voltage, V_{out}, for a given resonator current is given by:

$$V_{out} = I_s Z_1 \left(1 + \frac{Z_2}{Z_3}\right). \tag{5.61}$$

For both oscillators to be designed it will be found that the ratio of Z_2 and Z_3 is much less than one and therefore the output voltage can be approximated by:

$$V_{out} = I_s Z_1. \tag{5.62}$$

The oscillators are to be integrated in the bipolar DIMES01 process and to operate at a minimal supply voltage.

5.7.2 The non-tapped-resonator oscillator

5.7.2.1 The ideal input impedance

The input impedance of the non-tapped-resonator oscillator has to be -6 Ω. However, to ensure proper startup the input impedance is chosen to be -12 Ω at the cost of a worsening of the CNR by 3 dB [4] in the case of an instantaneous control.

5.7.2.2 Signal power

To keep the output signal of the oscillator within certain bounds, a limiter is chosen as the controlling block as it is more easy to implement than a time-averaged control. The most straightforward limiter that can be realized in a low-voltage application is the differential pair. Its transfer is from input voltage to output current and its limiting levels are on the order of 100 mV separated from each other. The only place to use this limiter is at the place of Z_3. The output voltage will therefore be on the order of 100 mV. To obtain a maximum resonator current of about 10 μA, Z_1 has to be:

- $Z_1 = 10\ \mathrm{k}\Omega$.

To obtain the intended input impedance, the ratio of the other two impedances has to be:

- $\frac{Z_2}{Z_3} = 0.0012$.

As the transfer of the limiter will be real, a resistor is chosen for Z_2.

5.7.2.3 Noise

For the input stage of the active part, a CE stage is chosen. To find the optimum collector current, i.e. for which the noise level is minimal, all the noise sources are transformed to an equivalent input noise voltage. The equivalent noise-power-density spectrum, S_v, is given by:

$$S_v = 4kTR_s + 4kTR_t\left(1+\frac{R_s}{R_1}\right)^2 + \frac{4kT}{R_1}R_s^2 + 4kT\left(\frac{1}{2}r_e + r_b\right) \times \left(1+\frac{R_s}{R_1}\right)^2 + 2qI_B\left[R_s + R_t\left(1+\frac{R_s}{R_1}\right)\right]^2 \qquad (5.63)$$

where R_t is the parallel connection of R_2 and R_3, $r_e = \frac{V_T}{I_C}$, V_T is the thermal voltage and I_B and I_C are the base and the collector current of the transistor, respectively. It was assumed that the two equivalent noise sources at the input of the bipolar transistor are given by:

$$S_{vbip} = 4kT\left(\frac{1}{2}r_e + r_b\right), \qquad (5.64)$$

$$S_{ibip} = 2qI_B. \qquad (5.65)$$

Minimizing expression (5.63) with respect to the collector current yields the following for the optimum collector current, $I_{C_{opt}}$:

$$I_{C_{opt}} = \frac{V_T\sqrt{\beta}}{\frac{R_1R_s}{R_1+R_s} + R_t}, \qquad (5.66)$$

where β is the current-gain factor of the transistor. In most practical cases $R_t \approx R_2$ and $R_s \ll R_1$, as a result, this equation can be simplified to:

$$I_{C_{opt}} = \frac{V_T\sqrt{\beta}}{R_s + R_2}. \tag{5.67}$$

At this point some conclusions can be drawn. From equation (5.63) it follows that for a relatively low noise contribution of the active part to the phase noise it must hold that:

- $R_2 \ll R_s$;
- $r_b \ll R_s$;
- $r_e \ll R_s$.

As R_s is only $6\ \Omega$, these constraints are not very easy to reach. For R_2 a value of:

- $R_2 = 10\ \Omega$,

is about the lowest practical resistance which can be chosen. Consequently, for R_3 it holds that:

- $R_3 \approx 8300\ \Omega$.

To obtain a low base resistance for the input transistor, a transistor is chosen with ten base contacts resulting in a base resistance of about:

- $r_b = 60\ \Omega$.

Lower base resistances are possible, but larger transistors are then required.

From equation (5.67) follows the optimum collector:

$$I_{C_{opt}} = 16\ \text{mA}, \tag{5.68}$$

for $\beta = 100$. For this collector current, the equivalent input noise resistance of the transistor is about $1.5\ \Omega$, which is much lower than the noise already introduced by the base resistance. Therefore, to save power the collector current is reduced which inherently results in a reduction of the loop gain. The lowest acceptable collector current equals about:

- $I_{C_{opt}} = 200\ \mu\text{A}$.

The equivalent input noise resistance is now approximately $65\ \Omega$ which is about the same as the noise resistance already present.

The total equivalent input noise voltage of this oscillator is now given by:

$$S_{veq} = 4kT\left(R_s + R_2 + \frac{1}{2}r_e + r_b\right) \approx 4kT \cdot 141\ \Omega. \tag{5.69}$$

Thus, the oscillator constant is worsened by about 14 dB.

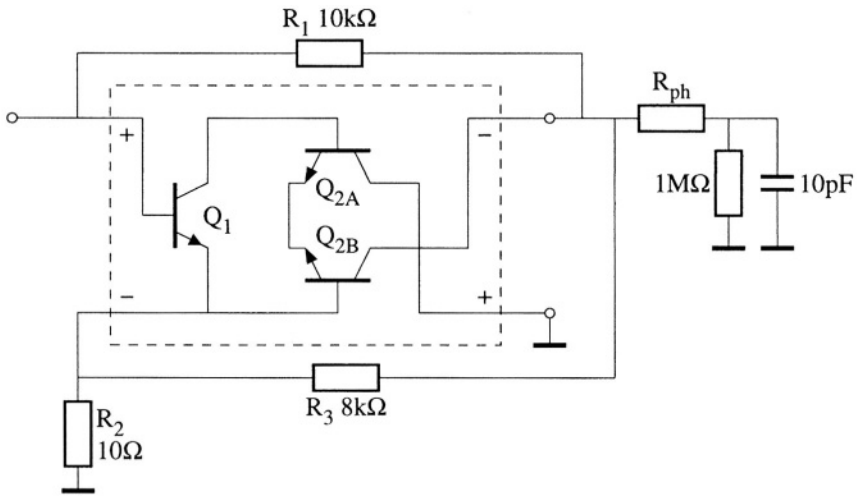

Figure 5.24: A two-stage implementation of the nullor.

5.7.2.4 Bandwidth

To obtain enough bandwidth, the nullor had to be implemented by two CE-stages, see figure 5.24. For both loops, the one with open input and the one with shorted input, the poles are determined by:

- the load capacitance and R_3;
- the input impedance of the differential pair and the substrate capacitance of Q_1.

The big difference, however, is found in the DC loop gain. These are: $A\beta_{op} = -2200$ and $A\beta_{sc} = +3.5$. This big difference is a result of the attenuation in $A\beta_{sc}$ caused by the very low value of R_2 compared with $r_{\pi 1}$. The resulting zeros in the input impedance are found at +2 MHz and -6 MHz (the roots of $1 - A\beta_{sc}$ become zeros of Z_{in}, see equation (5.58)). The zero at +2 MHz results in a phase of about 20 degrees, which gives a reduction in the quality factor of about 15 %, see equation (5.51), which is acceptable. The relatively low loop gain also causes a deviation of the intended input impedance. This can be counteracted by slightly changing R_1 or R_2. As a result of its relatively low DC loop gain, the loop $A\beta_{sc}$ requires no additional frequency compensations.

The other loop is frequency compensated by means of a phantom zero realized by means of R_{ph}, which has to be chosen at 500 Ω. The resulting poles are placed in Butterworth at a frequency of about 30 MHz.

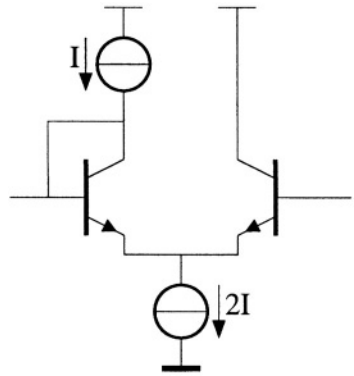

Figure 5.25: The limiter implemented by a differential pair.

5.7.2.5 The limiter

The limiter is implemented by the differential pair as depicted in figure 5.25. The gain from the input voltage, v_i, to the output current, i_o, of the differential pair in its linear region, is given by:

$$\frac{i_o}{v_i} = 0.5 g_m, \tag{5.70}$$

where g_m is the transconductance of one transistor. Thus, to obtain a transfer of $1/(8\ \mathrm{k}\Omega)$, a tail current of about $7\ \mu\mathrm{A}$ is required.

5.7.2.6 The total circuit

In figure 5.26 the total circuit is depicted. Transistor Q_{L1} and resistor R_{L1} are used as a level shift in order to prevent the tail-current source of the output stage from saturating. Transistor Q_{L2} is also a level shift.

The implementation of the required current sources is done by means of current mirrors. The current source feeding the input stage requires series feedback in order to increase its output impedance to prevent it from reducing the loop gain $A\beta_{sc}$. As a result of this series feedback, the noise level of the current sources is also reduced. The minimum required supply voltage is 1.3 V and the total current consumption is about $800\mu\mathrm{A}$.

From simulation, the equivalent input noise voltage of the complete circuit is found to be:

$$S_{veq} = 4kT \cdot 220\Omega. \tag{5.71}$$

The noise level is again increased. This is due to the lower bias current of the input stage ($180\ \mu\mathrm{A}$) which is caused by errors in the mirrors, and the contribution of the bias sources, which is still considerable.

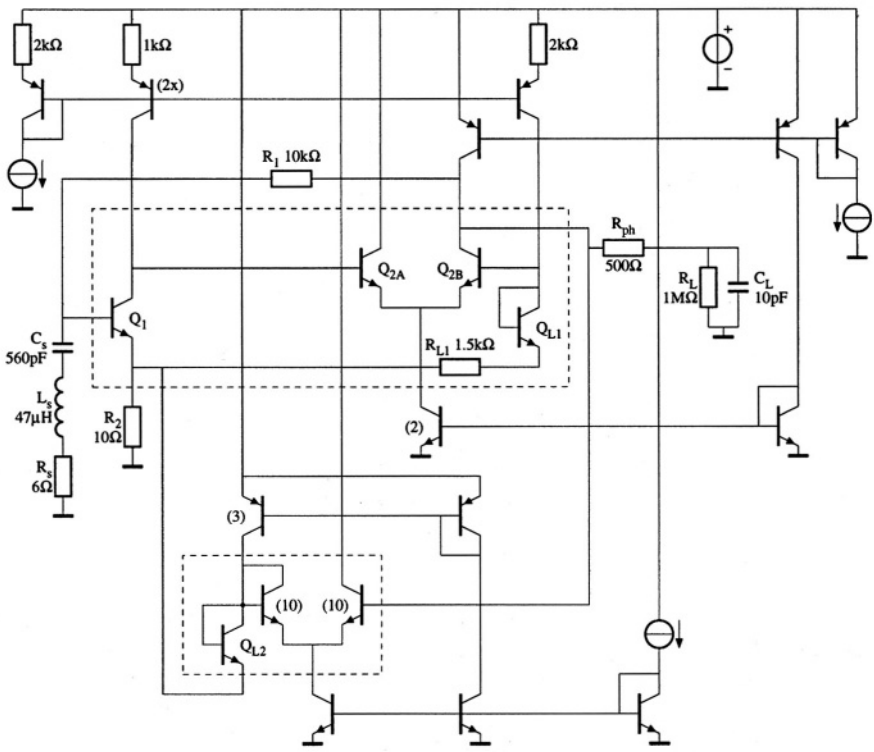

Figure 5.26: The complete schematic of the non-tapped-resonator oscillator.

Thus this oscillator is about 16 dB away from the maximum CNR.

5.7.3 The tapped-resonator oscillator

In this section the design of a tapped-resonator oscillator is discussed. The first thing to be done now is to derive the optimum tap factor. Therefore equation (5.63) is rewritten with R_s replaced by $t_f^2 R_s$ to account for the impedance transformation resulting from the tapping:

$$S_v = 4kTt_f^2R_s + 4kTR_t\left(1+\frac{t_f^2R_s}{R_1}\right)^2 + \frac{4kT}{R_1}t_f^2R_s^2 + 4kT\left(\frac{1}{2}r_e + r_b\right) \times \left(1+\frac{t_f^2R_s}{R_1}\right)^2 + 2qI_B\left[t_f^2R_s + R_t\left(1+\frac{t_f^2R_s}{R_1}\right)\right]^2. \quad (5.72)$$

As the resonator impedance is changed, the optimum collector current will also change. It is given by:

$$I_{C_{opt}} = \frac{V_T\sqrt{\beta}}{\frac{t_f^2R_sR_1}{t_f^2R_s+R_1} + R_2}. \quad (5.73)$$

In Section 5.5.1 it was shown that the output current of the resonator reduces proportionally with an increasing tap factor. Therefore resistor R_1 has to be made proportional to the tap factor in order to remain at the same voltage level at the output of the oscillator. Thus:

$$R_1 = t_f \cdot R_{1_0}, \quad (5.74)$$

where R_{1_0} is the value of R_1 without tapping; $R_{1_0} = 10\ \text{k}\Omega$ for this example.

As the resonator voltage also depends on the tap factor (the output current is assumed to be given), the optimum tap factor has to be found from the maximization of the CNR and not from the minimization of the equivalent input noise voltage. Thus the function to be maximized is:

$$CNR \propto \frac{V_{res_{in}}^2(t_f)}{S_{v_{eq}}(t_f, I_{C_{opt}})} = \frac{t_f^2V_s^2}{S_{v_{eq}}(t_f, I_{C_{opt}})}, \quad (5.75)$$

where $V_{res_{in}}$ is the input voltage of the tapped resonator and V_s is the input voltage of the non-tapped resonator, which is independent of the tap factor.

This expression, normalized to the maximum CNR, is plotted in figure 5.27. As clearly follows from the figure, there is an optimum for which the CNR is only 15 % away from the maximum attainable CNR.

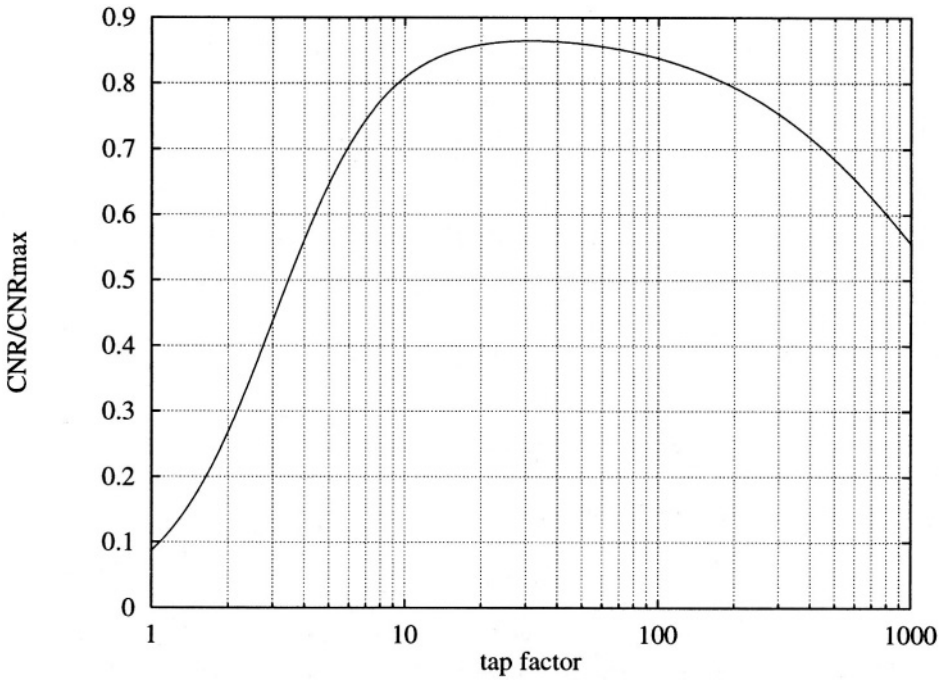

Figure 5.27: The CNR normalized to the maximum attainable CNR as a function of the tap factor.

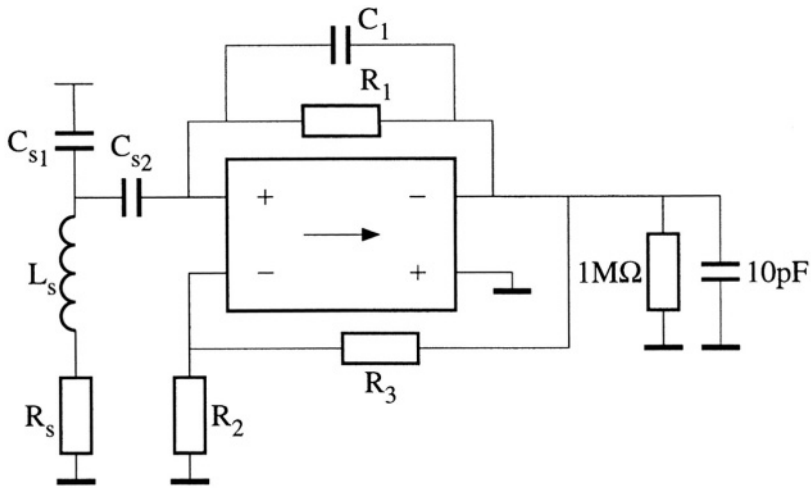

Figure 5.28: The configuration of the tapped-resonator oscillator.

From maximizing equation (5.75) the following expression is found for the optimal tap factor, $t_{f_{opt}}$:

$$t_{f_{opt}} = \sqrt[3]{\frac{2(R_2 + r_b)R_{1_0}}{R_s^2}}. \tag{5.76}$$

This equation holds when the following constraints are fulfilled: $R_s \cdot R_2 \ll R_{1_0}^2$, $R_2 \ll R_3$, and $R_s \cdot r_b \cdot \sqrt{\beta} \ll R_{1_0}^2$.

In contrast to the optimum tap factor which was found in Section 5.5.3, the optimum tap factor found here is a single value and not a range of values. This is because the noise power of resistor R_1 reduces proportionally to only t_f, see the dotted line in figure 5.13. But, as the noise contribution of R_1 is relatively small, the maximum is relatively flat, see figure 5.27.

For this example the optimum tap factor is found to be:

$$t_{f_{opt}} = 34. \tag{5.77}$$

This optimum tap factor is close to the quality factor of the resonator (Q_s=48) and will result in a reduction of the effective Q to about 33 and the CNR would be halved. As the maximum is relatively flat, the t_f is halved and the effective Q is now about 43, with a corresponding CNR degradation of 20 %. The power density of the equivalent noise voltage increases by only 1 %.

The tapped resonator and the active part are depicted in figure 5.28. The following values are used for the resonator capacitors (the inductor is kept the same):

- $C_{s1} = 560$ pF;
- $C_{s2} = 39$ pF;

resulting in a resonance frequency equal to:

- $f_{res} = 949$ kHz.

Capacitor C_1 is used for compensating for the equivalent parallel capacitance of the tapped resonator. Its value is derived in Section 5.7.3.3.

The actual tap factor can be calculated from the two capacitances, which equals:

$$t_f \approx 15. \tag{5.78}$$

The resonator resistance at resonance is therefore:

$$t_f^2 R_s = 1350\Omega. \tag{5.79}$$

5.7.3.1 Signal power

In order to have the same oscillator output voltage as the previous design (for the same inductor current), R_1 has to be equal to:

$$R_1 = t_f \cdot R_{1_0} = 150 \text{ k}\Omega, \tag{5.80}$$

and thus the ratio of R_2 and R_3 to obtain the required input impedance has to be equal to:

$$\frac{R_2}{R_3} = 0.009 \times 2 = 0.018, \tag{5.81}$$

in which the factor 2 accounts for the double input impedance to ensure startup.

5.7.3.2 Noise

The optimum collector current for the input stage follows from equation (5.73), which yields:

$$I_{C_{opt}} = 190 \ \mu\text{A}. \tag{5.82}$$

However, as the noise contribution of the active part is relatively small compared with the noise contributed by the resonator (the actual CNR is close to the maximum CNR), the collector bias current can be reduced in order to save power. A current of about 100 μA showed to be convenient as it proved to give an acceptable HF behavior. At this collector current the CNR ratio is reduced by only 3 %.

The equivalent noise resistance calculated at R_s and *including* R_s, equals:

$$R_{eq} \approx R_s + \frac{R_2 + r_b + \frac{1}{2}r_e}{t_f^2} \approx 7\Omega, \tag{5.83}$$

which is only slightly larger than R_s. In this equation R_2 is assumed to be 10 Ω.

5.7.3.3 Bandwidth

Capacitor C_1 in figure 5.28 is required to cancel the equivalent parallel capacitance of the tapped resonator, which is 36 pF (560 pF in series with 39 pF). To compensate exactly for this capacitance, C_1 has to be 0.3 pF. But, as the input impedance of the oscillator is to be designed with a double magnitude, C_1 also has to be twice as large, in order to cancel the capacitance at startup. The second stage of the active part is implemented by a differential pair, as it was for the previous design.

Again, $A\beta_{op}$ and $A\beta_{sc}$ are studied for the frequency behavior of the input impedance. The loop, for which the input is left open, has a loop gain of about -24. This is considerably reduced in comparison to the previous oscillator. The main cause is the lower value of R_3 and the higher value of R_1. The loop consists of a phantom zero due to C_1 and R_1 (-2.1 MHz). This phantom zero is equivalent to a pole in the input impedance (for infinite loop gain), compensating for the parallel capacitance of the tapped resonator. Besides the phantom zero two poles are also found in the loop, one at -90 kHz and one at -29 MHz. These two poles are from the input and output stage, respectively, and they are split by the base-collector capacitance of Q_1. As a result the closed loop pole is found at -1.2 MHz instead of the intended -2.1 MHz. The extra phase shift of the input impedance at the resonance frequency is about -30 degrees. This is acceptable for the moment [5].

The positive loop, for which the input port is shorted, has a loop gain of about 4. Only the pole from Q_2 (-2.7M) is relevant, resulting in the input impedance in a zero at +6.4 MHz. It should be noted that by shorting the input, the local feedback caused by the base-collector capacitance is broken, and thus the poles of Q_1 and Q_2 are not split.

As will be seen in the next section, for implementing the limiter it is power efficient to have a higher value for R_3, consequently R_2 will also become higher. The final values for R_2 and R_3 are respectively, 20 Ω and 1 kΩ.

5.7.3.4 The limiter

For the limiter, the same configuration is used as for the previous example. When a transconductance in its linear region of $1/(500\ \Omega)$ is realized (which was the first choice), the tail current of the differential pair, see figure 5.25, should be 200 μA. Therefore the value of R_2 and the transconductance of the limiter in its linear region are doubled. For the noise this gives a negligible small increase whereas it is more advantageous for the bandwidth of the input impedance. Now the pole in the transfer comes closer to the phantom zero, and as a result the additional phase is reduced to -20 degrees.

[5] To reduce the effect of the base-collector capacitance to a large extent, a current buffer could be used preceding the stage [14].

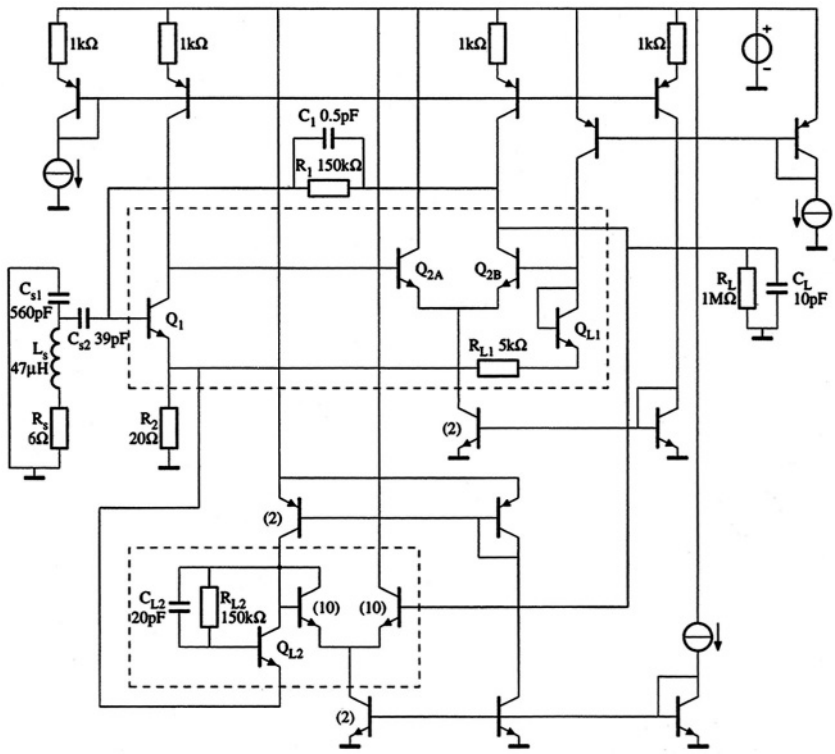

Figure 5.29: The circuit diagram of the tapped-resonator oscillator.

5.7.3.5 The total circuit

In figure 5.29 the total schematic of the oscillator is depicted. This circuit is also supplied from a 1.3 V supply voltage in order to be able to also use the local series feedback in the current sources. The oscillator consumes about 800 μA. The circuit diagram closely resembles that of the other oscillator. A difference can be found at the level shift required for proper functioning of the limiter. It also uses a diode-connected transistor Q_{L2} but now a resistor, R_{L2}, is placed in series with the base terminal. This is necessary for compensating at the limiter input for the voltage drop which is found across R_1 and caused by the base current of Q_1. Capacitor C_{L2} is required to obtain the correct limiter transconductance again at the frequency of oscillation.

From small-signal simulations the equivalent input noise voltage at R_s and including R_s still equals:

$$S_{veq} \approx 4kT \cdot 7\Omega. \tag{5.84}$$

Therefore this oscillator reaches the maximum CNR within 0.7 dB (the 3 dB

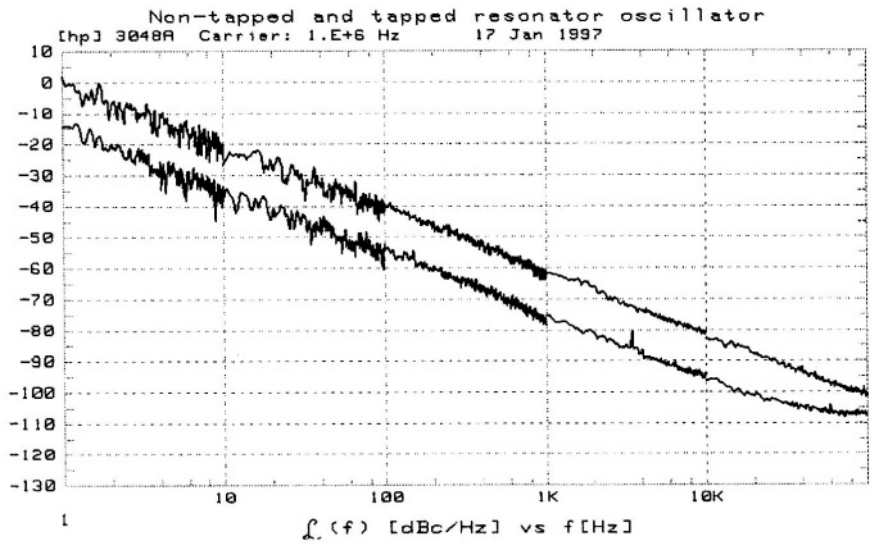

Figure 5.30: The measured $\mathcal{L}(\Delta\omega)$ of the two oscillators. The upper and lower graph corresponds to the non-tapped and tapped-resonator oscillator, respectively.

reduction due to the excess loop gain of the oscillator is ignored here for the moment), which is much better than the other oscillator. Of course, due to the additional phase shifts a lower effective Q is found, which is however of the same order of magnitude as that of the other example. The extra degradation of the CNR may be expected due to the Q-degradation caused by the tapping, which is about 1 dB.

5.7.4 Measurement results

The two oscillators are integrated and their performance was measured on the HP phase-noise measurement system (HP3048A). The results are depicted in figure 5.30. Both oscillators were measured under the same conditions ($V_{SUP} = 1.3$ V and for both oscillator $I_{SUP} \approx 0.8$ mA). For the tapped-resonator oscillator, an amplitude of 100 mV resulted in relatively high power levels of the harmonics. Therefore, the amplitude was reduced to 50 mV. Consequently, the minimum oscillator constant increases to:

- $OC_{min} = -140$ dBc/Hz.

From the measurements, the $\mathcal{L}$ of the non-tapped and tapped-resonator oscillator at $\Delta f = 10$ kHz were found to be, respectively:

- $\mathcal{L}_{non-tapped}(\Delta f = 1 \text{ kHz}) \approx -61$ dBc/Hz;
- $\mathcal{L}_{tapped}(\Delta f = 1 \text{ kHz}) \approx -75$ dBc/Hz.

These two differ by 14 dB, which perfectly agrees with the calculations and simulations. The corresponding oscillator constants are given by:

- $OC_{non-tapped} = -121$ dBc/Hz;
- $OC_{tapped} = -135$ dBc/Hz.

These figures are 5 dB to high when compared with the expected -140 dBc/Hz. However, the oscillator constant of -140 dBc/Hz was found assuming an intrinsic Q (no Q-degradation), infinite bandwidth of the active part and no increase of the phase noise as a result of limiting in the feedback network. The corresponding degradation is given by:

- $\Delta\mathcal{L}_{Q-degradation} \approx 0.8$ dB ($t_f = 15$ and $Q_s = 48$);
- $\Delta\mathcal{L}_{finite-bandwidth} \approx 1$ dB ($\phi \approx -20^o$);
- $\Delta\mathcal{L}_{limiter} \approx 3$ dB for an excess loop gain of 2.

These additional contribution amounts to about 5 dB which agrees with the 5 dB higher measured $\mathcal{L}$.

With respect to the figure two additional comments can be made. First, for the tapped-resonator oscillator the noise floor can be found starting at about 60 kHz from the carrier whereas it is for the non-tapped-resonator oscillator beyond a distance of 100 kHz. This results from, besides the lower phase noise of the former oscillator, the fact that the collector current of the input stage of the former oscillator was chosen about the half of the optimum value (100 μA instead of 190 μA). Consequently, the equivalent voltage noise of the active part is doubled. This source is one of the causes of the noise floor, see section 5.5.5. Second, the measurement results do not shown any contribution of 1/f noise. Consequently, the transistors in the active part do have a very low excess-noise corner frequency, below 1 Hz!

5.8 Conclusions

In this chapter the structured design of tapped harmonic oscillators was dealt with. In order to obtain a maximal signal power in the resonator relative to the noise level for a given supply voltage, the series resonator was shown to be the best choice. This resonator has a lower impedance than the corresponding parallel resonator.

The CNR of the oscillator is limited by the intrinsic quality factor and noise level of the resonator. When compensating for the losses of the resonator, in

order to get it oscillating, by means of an active circuit, the equivalent voltage noise of this active part proved to be a problem due to the relatively low impedance of the resonator.

Tapping the resonator has been shown to be a very convenient method for obtaining a better noise match. As a result of tapping, the impedance level of the resonator is increased without degrading the intrinsic CNR. The noise match is then found for lower bias currents of the input stage, which is more power efficient, and the relative contribution of the active part to the noise of the oscillator is considerably reduced.

For an increasing tap factor, the coupling of the resonator and active part weakens. Therefore, for relatively high tap factors it may be expected that the quality factor of the total oscillator reduces with respect to the intrinsic quality factor of the resonator. It was shown that for a tap factor that is relatively low when compared with the quality factor, this degradation can be ignored and then the noise minimization can be done assuming a constant quality factor.

Tapping inherently introduces a capacitance parallel to the resonator. As the bandwidth of the active part, i.e. the active negative impedance, is relatively high, parasitic oscillations may occur. This can be solved by not only undamping the resonator by a negative resistance but also by a negative capacitance.

The effectiveness of tapping is illustrated by a design example. For this example the calculated increase of the CNR is about 14 dB which also followed from the measurements.

Bibliography

[1] E.A. Vittoz, M.G.R. Degrauwe, and S. Bitz. High-performance crystal oscillator circuits: Theory and application. *IEEE Journal of Solid-State Circuits,* 23(3):774–783, June 1988.

[2] C.J.M. Verhoeven, J.R. Westra, A. van Staveren, and A.H.M. van Roermund. Low noise oscillators. In J.H. Huijsing et al., editors, *Analog Circuit Design,* pages 65–85. Kluwer, Dordrecht, 1995.

[3] W.A. Edson. *Vacuum-Tube Oscillators.* John Wiley & Sons, New York, 1953.

[4] C.A.M. Boon. *Design of High-Performance Negative-Feedback Oscillators.* PhD thesis, Delft University of Technology, September 1989.

[5] D.B. Leeson. A simple model of feedback oscillator noise spectrum. *Proceedings of the IEEE,* 52(2):329–330, 1966.

[6] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[7] H. Lindenmeier. Noise matching techniques in transistor oscillators. In *Proceedings IEEE Symposium on Circuits and Systems, Houston,* pages 1052–1055, April 1980.

[8] G. Braun and H. Lindenmeier. Transistor oscillators with impedance noise matching. *IEEE Transactions on Microwave Theory and Techniques,* 39(9):1602–1610, September 1991.

[9] J. Craninckx and M. Steyaert. Low-noise voltage-controlled oscillators using enhanced LC tanks. *IEEE Transactions on Circuits and Systems II,* 42(12):794–804, December 1995.

[10] M. v.d. Gevel. private communication.

[11] M.J. Underhill. Oscillator noise limitations. In *IERE Proceedings of the Conference on Electromagnetic Compatibility,* volume 39, pages 109–118, 1978.

[12] E.H. Nordholt and C.A.M. Boon. A systematic approach to the design of single-pin integrated crystal oscillators. In *Proceedings of the 30th Midwest Symposium on Circuits and Systems,* pages 753–756, 1988.

[13] R.B. Blackman. Effect of feedback on impedance. *The Bell System Technical Journal,* 22(3):269–277, October 1943.

[14] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

Chapter 6

Bandgap references

6.1 Introduction

For many years bandgap references have been used as voltage references in various fields of application; for instance, in DA converters [1]. Depending on the digital input word, the analog output voltage is a fraction of the internal reference voltage. As for many applications this digital-to-analog conversion is not allowed to be temperature dependent, the reference voltage has to be temperature independent. Nowadays, the resolution of DA converters is very high and consequently the reference voltage must be very stable as each variation in the reference voltage is directly seen in the DA-converter output.

In the last decade the automotive industry has been using more and more electronics to realize more functions and larger systems in cars. The automotive environment, however, is very extreme; the supply voltage can have very large voltage transients (i.e. 80V for 12V systems [2]) and temperature variations can be in the range of -15°C to 105°C. Under these circumstances the electronics still need to work correctly. A supply-voltage regulator handles the stabilization of the supply voltage in the car. The reference for the regulator may be a bandgap reference [2]. This bandgap reference does not need to be very accurate but it must be very robust. It must be able to withstand the extreme automotive environment.

An equivalent kind of application is the bandgap reference used as reference voltage in battery-operated DRAMs [3]. The bandgap reference is also used as a reference for the power-supply-voltage stabilization, but now the power consumption is of prime importance.

All of the bandgap references used in this diverse variety of applications are based on the idea of Hilbiber in 1964 [4]. After his publication a vast number of articles appeared describing other topologies and components being used. In this

chapter a structured design method is presented which structures the knowledge gained in the past decades concerning bandgap reference design and from that an extrapolation is made. Therefore, first a historical overview is given to elucidate the evolution of bandgap reference design over the years. Subsequently, the structured design method is presented. To get a clear insight into the bandgap reference to be synthesized, the basic function is studied, including a general description of the temperature compensation and an inventory of the key parameters. Then, as the basic structure of the bandgap reference is found, the relation to the fundamental design aspects is discussed which is followed by a more detailed discussion of them, i.e. noise behavior, bandwidth capability and signal power. Thereafter, some special structures for the bandgap reference, resulting from reductions at the mathematical level are given, which is followed by two design examples. Finally, conclusions are given.

6.2 Historical overview

In 1964, Hilbiber published the first bandgap reference [4]. He proposed to compensate for the temperature behavior of a base-emitter voltage by adding and subtracting several base-emitter voltages with different first-order temperature behaviors. As his method used several stacked base-emitter voltages, the required power-supply voltage was relatively large compared with the reference voltage. In 1971 Widlar proposed a new basic scheme of the bandgap reference [5] requiring a lower supply voltage and this subsequently became commonly used. His method was based on the compensation of the first-order temperature behavior of the base-emitter voltage with a voltage which is proportional to the absolute temperature (PTAT). He had found this PTAT voltage in 1965 [6] by using the difference of two junction voltages. Seen mathematically, this method is identical to the Hilbiber's method, only the implementation was different. Where Hilbiber first made two appropriate stacks of base-emitter voltages, with a different first-order temperature behavior, and subsequently used the difference as the first-order temperature-compensated reference voltage, Widlar first made a relatively small voltage with a linear temperature behavior (PTAT), whereupon this voltage was amplified to cancel the first-order term of a base-emitter voltage. Widlar implemented the amplification of the voltages closer to the output of the reference.

In 1973, Kuijk made an integrated bandgap reference [7] using Hilbiber's idea [4]. He used, however, an additional scaling factor in his reference such that other output voltages, other than the bandgap voltage, could also be realized.

Widlar's method uses at least three bipolar transistors, two for the PTAT voltage and one for the base-emitter voltage. Hilbiber's method uses a stack of n and m transistors with n and m usually larger than 2. Brokaw published, in 1974 [8], a bandgap reference which, more or less, can be seen as the combination

of Hilbiber's and Widlar's methods. He uses two "stacks" of one transistor to realize a PTAT voltage (Widlar) and amplifies this. To this PTAT voltage a base-emitter voltage from a transistor from one of the stacks (Hilbiber) is added to realize the bandgap voltage. Consequently, the minimum number of transistors required was reduced to two.

Meijer proposed in 1976 [9] to realize the amplifying action for the PTAT voltage in a different way than was done before.

All the bandgap references until then were only first-order compensated. The remaining temperature dependency was mainly of the second order and given by the behavior of the bandgap voltage and the physical phenomena underlying the base-emitter voltage. Thus, as these behaviors and phenomena are equal for all the proposed structures, the remaining temperature dependency is largely the same for all the references. However, some differences may exist due to the way off implementation. For one implementation, the base currents may have a negative effect on the temperature behavior, as for another implementation some compensation of the second-order temperature behavior may be obtained (but not intended).

In 1978, Widlar proposed a configuration which exhibits an intended compensation of the second-order temperature behavior [10], [11]. The principal difference with previous references is that the transistors generating the base-emitter voltages are biased at collector currents having different temperature dependencies. This results in different second-order temperature behaviors for the two base-emitter voltages and (partial) compensation became possible. In 1981, Dobkin [12] presented a circuit also with the correction of the parabolic curvature, i.e. a curvature corrected bandgap reference. He used Widlar's idea [5] as the basis for linear compensation. For the curvature correction he made the current ratio in his PTAT cell temperature dependent by using currents which were PTAT and CTAT (complementary to the absolute temperature).

The basic linear compensation of Hilbiber [4] was extended with curvature correction by Meijer in 1982 [13]. The curvature correction was realized by biasing one of the two stacks with a PTAT current and the other stack with a constant current derived from the output voltage of the bandgap reference.

A principally different way of curvature correction was introduced by Lee [14]. Instead of using different temperature behaviors for the collector bias currents to realize the curvature correction, in [14] the fact that the base current has an exponential temperature behavior is used. He shows with simulations that by using this method, better results can in principle be obtained than with the previous methods (approximately a factor 2).

In the late seventies, when weak-inversion MOS transistors came more into the picture [15], bandgap references also became implemented in CMOS processes. In 1979, both Vittoz [16] and Tzanateas [17], published a CMOS bandgap reference. The PTAT voltage, required for the linear temperature compensa-

tion, was realized by using weak-inversion MOS transistors. The transistor for generating the base-emitter voltage was made by a bipolar substrate transistor, having the drawback of an inherently grounded collector, such that the collector current has to be set via the emitter current and consequently introduces an inaccuracy.

The bandgap reference presented in [18] uses only MOS transistors for the amplifiers. The PTAT and base-emitter voltage are realized by bipolar substrate transistors. Different temperature behaviors for the collector currents are used to realize a curvature correction. The same was done by Lin [19] only he used a different model for the base-emitter voltage.

Salminen used in his bandgap reference, reported in 1992 [20], the non-linear temperature behavior of a diode-connected MOS transistor. With this non-linear voltage the curvature of a conventional first-order compensated bandgap reference was reduced.

As may be clear from this history of bandgap references, the basic idea of all these references is the same, only implementation details vary, i.e. mathematical operations are implemented in a different manner and/or different order. Thus, it should be possible to find a general description of the design of bandgap references covering all these designs.

6.3 The basic function

In the introduction it was shown that the bandgap reference is a widely used basic building block for generating reference voltages. For the ideal voltage reference, ideal specifications can be defined. These specifications also hold for the ideal bandgap reference.

The ideal bandgap reference has an output impedance equal to zero such that any load current does not alter the reference voltage. Further, it may not make a difference under what environmental circumstances the output voltage of the bandgap reference is measured; it should always have the same value. The environmental circumstances can be temperature, humidity, et cetera. Thus, the bandgap reference may not be sensitive or dependent on any of these parameters. Of course, the reference voltage has to be constant over time as well. This means that the output voltage of the ideal bandgap reference is noise free; no time dependent variation are to be found at the output.

To be able to achieve this ideal source as closely as possible with a bandgap reference implementation, a formal description of the bandgap reference is of prime importance. The starting point is the basic function to be realized: a reference voltage, V_{REF}, related to the bandgap voltage at 0K, $V_G(0)$ or:

$$V_{REF} = x \cdot \frac{E_G(0)}{q} = x \cdot V_G(0), \tag{6.1}$$

where q is the electron charge (1.6e-19 C) and x is a constant scaling factor accounting for reference voltages not equal to the bandgap voltage. For realizing a bandgap reference, at least one component must be available of which a port voltage is related to the bandgap energy. Generally, this core component of the bandgap reference is the bipolar transistor as the base-emitter voltage is related to the bandgap energy in a very simple and accurate way, see Section 6.3.1. Even in most MOS bandgap references at least one bipolar transistor/junction is used for the relation to the bandgap energy [14], [16]-[20], [21] and [22]. However, in [23] the difference of two gate-source voltages with different polysilicon-gate work functions is proposed to realize a voltage which is related to the bandgap energy. But, since with bipolar transistors as core component more-accurate bandgap references can be made than with MOS transistors, because of better matching [24] and modeling, and thus reaching the ideal reference the best, here the bipolar transistor is chosen as the core element of the bandgap reference. Still, MOS transistors can be favorably used in other parts of the bandgap reference because of their high DC gain.

6.3.1 The base-emitter voltage

When the bipolar transistor is chosen as the core component of the bandgap reference, the base-emitter voltage is the voltage which is related to the bandgap energy and has to be used. In this section the relation between the base-emitter voltage and the bandgap energy is discussed [25], [26]. The basic relation describing the collector current, I_C, as a function of the base-emitter voltage, V_{BE}, is given by:

$$I_C(T) = I_S(T)\left[\exp\left(\frac{qV_{BE}}{kT}\right) - 1\right], \tag{6.2}$$

where T is the absolute temperature, k is the Boltzmann constant (1.38e-23 J/K) and I_S is the saturation current. The '-1' term in the expression is ignored in the rest of this chapter as it is assumed that the injection current is much larger than the saturation current [27]. For I_S the following expression holds:

$$I_S(T) = \frac{qAn_i^2(T)\overline{D}(T)}{N_B}, \tag{6.3}$$

where A is the emitter area, $n_i(T)$ is the intrinsic carrier concentration, $\overline{D}(T)$ is the mean minority-carrier diffusion constant in the base and N_B is the Gummel number of the base region. The intrinsic carrier concentration is related to the bandgap energy according to:

$$n_i^2(T) = CT^3 \exp\left[-\frac{E_G(T)}{kT}\right], \tag{6.4}$$

where C is a constant and E_G is the bandgap energy as a function of the temperature. The temperature behavior of $\overline{D}$ is found from the Einstein relation:

$$\overline{D}(T) = \frac{kT}{q}\overline{\mu}(T), \tag{6.5}$$

where $\overline{\mu}(T)$ is the mean mobility of the minority carriers in the base region. The temperature behavior of this mean mobility can be defined by:

$$\overline{\mu}(T) = BT^{-n}, \tag{6.6}$$

where B and n are constants. Combining equations (6.2)-(6.6) the following expression is found:

$$I_C(T) = C'T^{\eta}\exp\left[\frac{qV_{BE} - E_G(T)}{kT}\right], \tag{6.7}$$

where $\eta = 4 - n$ and $C' = \frac{A \cdot B \cdot C \cdot k}{N_B}$ are constants. Rewriting this expression to make the base-emitter voltage explicit, yields:

$$V_{BE}(T) = V_G(T) + \frac{kT}{q}\ln\left[\frac{I_C(T)}{C'T^{\eta}}\right]. \tag{6.8}$$

According to this expression the base-emitter voltage is directly related to the bandgap energy. This expression can be further simplified when it is assumed that the currents that can be realized conveniently are the constant current, derived from a bandgap reference, and the PTAT current, derived from the difference of two junction voltages. Both currents can be expressed by:

$$I(T) = I(T_r)\left(\frac{T}{T_r}\right)^{\theta}, \tag{6.9}$$

where T_r is a reference temperature and θ is the order of the temperature dependency. For the constant current $\theta = 0$ and for the PTAT current $\theta = 1$. Using this expression for the collector current in equation (6.8) yields the following for the base-emitter voltage:

$$V_{BE}(T) = V_G(T) + \frac{kT}{q}\ln\left[\frac{I_C(T_r)}{C'T_r^{\eta}}\right] - \frac{kT}{q}(\eta - \theta)\ln\left(\frac{T}{T_r}\right). \tag{6.10}$$

Finally, the base-emitter voltage can be rewritten such that the constant C' is removed:

$$V_{BE}(T) = V_G(T) - V_G(T_r)\frac{T}{T_r} + V_{BE}(T_r)\frac{T}{T_r} - \frac{kT}{q}(\eta - \theta)\ln\left(\frac{T}{T_r}\right), \tag{6.11}$$

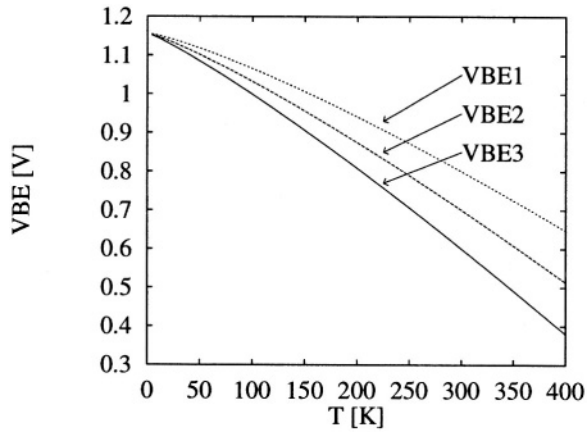

Figure 6.1: The base-emitter voltage as a function of the temperature for several values of V_{BE}, $V_{BE_1} > V_{BE_2} > V_{BE_3}$.

where $V_{BE}(T_r)$ and $V_G(T_r)$ are the base-emitter voltage and bandgap voltage at the reference temperature, respectively. This expression is depicted in figure 6.1 for several values of the base-emitter voltage. Some general conclusion can be drawn from this picture. Independent of the value of the base-emitter voltage at T_r, the base-emitter voltage at 0K always equals $V_G(0)$ and its temperature dependency is always negative. Further, for higher base-emitter voltages the temperature dependency of the base-emitter voltage reduces. How these properties can be used to realize a bandgap reference is the topic of the next section.

6.3.2 General temperature compensation

The expression for the base-emitter voltage, found in the previous section, is used as the core of the bandgap reference. Figure 6.2 depicts how a temperature independent voltage can be realized. The temperature dependency of the base-emitter voltage is compensated for by an additional voltage V_{COMP}. Consequently, the sum of V_{BE} and V_{COMP} equals, over the total temperature range, $V_G(0)$. As this voltage is independent of all kinds of parameters, a true reference is obtained. However, it should be noted that for higher-doped base regions, higher than 10^{17} cm^{-3}, bandgap narrowing may slightly change the value of $V_G(0)$ [28]. But still, the bandgap reference is obtained by compensating the base-emitter voltage with a complementary voltage.

For exact compensation V_{COMP} is a complex function of the temperature and not easy to realize. Therefore, the temperature dependency chosen for V_{COMP} is less complex and, consequently, a remaining temperature dependency

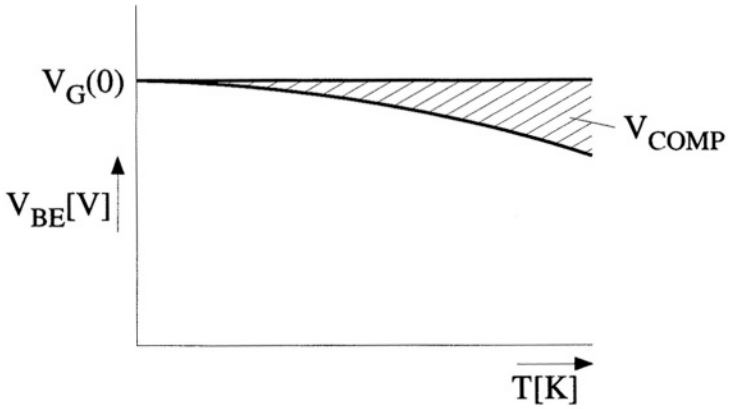

Figure 6.2: The principle realization of a bandgap reference.

is found. V_{COMP} can be realized with several temperature dependent voltages. For instance, in [14] V_{COMP} is partially realized by a voltage which is related to the temperature behavior of the base current. In that case, however, two different physical processes are used to compensate for each other. Their sensitivity to process variations, for instance, may be different resulting in a larger spread. In this book, base-emitter voltages are also used for the compensation voltage. Thus the bandgap reference is realized by a weighted sum of base-emitter voltages, i.e. a linear combination.

6.3.3 A linear combination of base-emitter voltages

With an appropriate linear combination of base-emitter voltages a bandgap reference can be realized. This linear combination should be chosen such that the required reference voltage is obtained and temperature compensation is performed. To be able to draw general conclusions the base-emitter voltage is rewritten as a Taylor polynomial around a reference temperature T_r. This reference temperature could be the center of the temperature range for which the bandgap reference is to be designed. First, the bandgap voltage is defined as:

$$V_G(T) = V_G(T_r) + \alpha_1(T - T_r) + \alpha_2(T - T_r)^2 + \cdots, \qquad (6.12)$$

where α_n are the Taylor coefficients of the bandgap voltage at the reference temperature. Defining the bandgap energy in this way makes the discussion independent of the particular description that is used for the bandgap voltage. Using this expression for the bandgap voltage, the Taylor polynomial of the

base-emitter voltage is found to be:

$$V_{BE}(T) = \underbrace{V_{BE}(T_r)}_{\text{Constant}} + \underbrace{\left[V_{BE}(T_r) - V_G(T_r) + \alpha_1 T_r - (\eta - \theta)\frac{kT_r}{q}\right]\frac{T - T_r}{T_r}}_{\text{First order}} + \underbrace{\sum_{n=2}^{\infty}\left[\alpha_n T_r^n - \frac{kT_r}{q}(\eta - \theta)\frac{(-1)^n}{n(n-1)}\right]\left(\frac{T - T_r}{T_r}\right)^n}_{\text{Second and higher order}}. \quad (6.13)$$

To alter the temperature behavior of a base-emitter voltage, only some of the variables appearing in this expression can be changed by the designer. They are given here in order of appearance:

- for the constant term, $V_{BE}(T_r)$;
- for the first-order term, $V_{BE}(T_r)$, θ;
- for the second and higher-order terms, θ.

Thus a designer can change the temperature behavior of a base-emitter voltage by choosing a different value for $V_{BE}(T_r)$ and/or by choosing a different order of temperature dependency for the collector bias current. The other parameters are given for a process, η, $V_G(T_r)$ and α_n, or are physical constants, k and q.

The reference voltage, V_{REF}, is made of a linear combination of i base-emitter voltages:

$$V_{REF} = \sum_{m=0}^{i} a_m V_{BE_m}(T), \quad (6.14)$$

where a_m are the scaling factors. In this way of describing the reference voltage, it is not made explicit which base-emitter voltage(s) form(s) the compensating voltage and which is the compensated voltage; each base-emitter voltage is treated equally. Introducing the Taylor polynomials in this expression results in the most general expression for the reference voltage:

$$V_{REF}(T) = \sum_{m=0}^{i} a_m V_{BE_m}(T_r) + \sum_{m=0}^{i} a_m \left[V_{BE_m}(T_r) - V_G(0)_1 - \frac{kT_r}{q}(\eta - \theta_m)\right]\frac{T - T_r}{T_r} + \sum_{m=0}^{i}\sum_{n=2}^{\infty} a_m \left[\alpha_n T_r^n - \frac{kT_r}{q}(\eta - \theta_m)\frac{(-1)^n}{n(n-1)}\right]\left(\frac{T - T_r}{T_r}\right)^n. \quad (6.15)$$

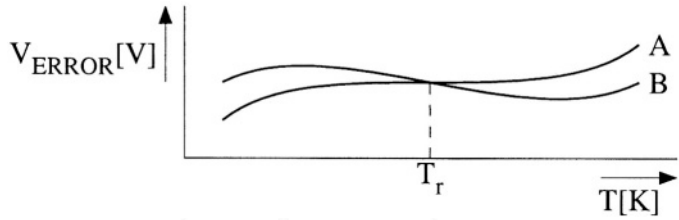

Figure 6.3: Two general ways of specifying the temperature dependency of a bandgap reference. A: maximum flat and B: lowest mean dependency.

In this expression $V_G(0)_1$ is the bandgap voltage at $0K$ extrapolated from $V_G(T_r)$ with a first-order polynomial: $V_G(0)_1 = V_G(T_r) - \alpha_1 T_r$. The reference voltage is described as a set of m equations of order ∞. Each base-emitter voltage introduces three degrees of freedom:

- $V_{BE_m}(T_r)$;
- θ_m;
- a_m.

6.3.3.1 Type of compensation

To obtain the ideal bandgap reference (with respect to temperature dependency) these $3 \cdot m$ parameters must be chosen such that the set of equations reduces to:

$$V_{REF}(T) = \sum_{m=0}^{i} a_m V_{BE_m}(T_r) + \sum_{n=1}^{\infty} 0 \cdot (T - T_r)^n. \qquad (6.16)$$

However, to obtain this, an infinite number of base-emitter voltages are required, i.e. $m = \infty$ (for the case that all the equations are independent). This is, of course, impractical. When less base-emitter voltages are used, i.e. the compensating voltage is not exactly complementary, a certain temperature dependency remains. The question is now how to use the available degrees of freedom such that this remaining temperature dependency is minimized. What specific method is used depends on the specification of the temperature dependency for the bandgap reference. On the one hand, it may be required that close to the reference temperature, T_r, the temperature dependency has to be minimal whereas the dependency at the boundaries of the temperature range may be less important. On the other hand a voltage range may be specified wherein the reference voltage must be for the whole temperature range. These two situations are depicted in figure 6.3. Curve A in figure 6.3 shows a temperature dependency which is minimized near T_r. This method corresponds to

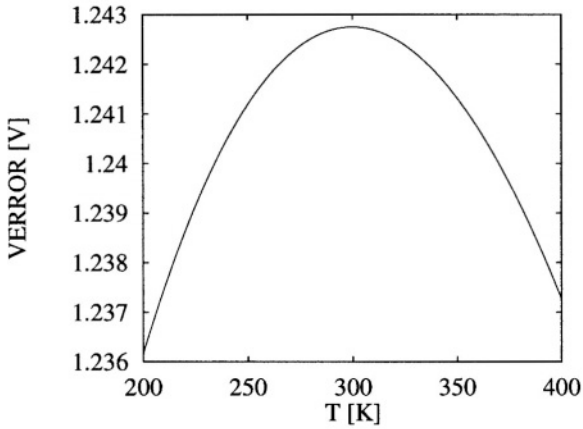

Figure 6.4: The remaining temperature dependency for a first-order compensated bandgap reference.

compensating as many orders as possible starting with the first-order temperature behavior and, depending on the number of degrees of freedom, the second order, third order, et cetera. For this temperature behavior the highest number of derivatives near T_r are zero and, consequently, the function is maximally flat.

Curve B in figure 6.3 depicts the reference voltage as a function of temperature when the reference voltage must remain within a specified voltage range. At T_r the temperature dependency is mostly not zero. This method of compensation is closely related to the first method. To make this more clear a bandgap reference is assumed to have two degrees of freedom, one to set the reference voltage and one for temperature compensation. When this second degree of freedom is used to compensate for the first-order temperature behavior, a remaining temperature dependency as depicted in figure 6.4 is obtained. To obtain a minimal error band, it must be examined to see whether or not some second-order dependency can be exchanged for some first-order dependency. Assume one scaling factor, a_x, is changed by da_x:

$$a_x \rightarrow a_x + da_x, \tag{6.17}$$

so that the second-order temperature behavior of the reference voltage a_{REF_2} is reduced by:

$$da_{REF_2} = da_x \cdot a_{VBE_2}, \tag{6.18}$$

where a_{VBE_2} is the second-order Taylor coefficient of the base-emitter voltage. Consequently, the first-order coefficient of the reference voltage, da_{REF_1}, changes by:

$$da_{REF_1} = da_x \cdot a_{VBE_1}, \tag{6.19}$$

in which a_{VBE_1} is the first-order Taylor coefficient of the base-emitter voltage. As the second-order behavior is even with respect to T_r both sides of the paraboloid in figure 6.4 change by:

$$dV_{REF} = da_x \cdot a_{VBE_2} \cdot \left(\frac{1}{2}\Delta T\right)^2, \tag{6.20}$$

where ΔT is the temperature range. In contrast, the first-order behavior is odd with respect to T_r and thus the variations in V_{REF} at the boundaries are opposite. Assuming that the maximum of the function does not change, a reduction in the total error is obtained when it holds that:

$$a_{VBE_1} \leq a_{VBE_2} \cdot \frac{1}{2}\Delta T. \tag{6.21}$$

As the ratio of the two Taylor coefficients is on the order of several thousands, the temperature range must be on the order of several thousands of Kelvin to profit from this exchange. For an exchange of second and third-order behavior, the argument holds. Thus, for the design of bandgap references it makes no sense to exchange between an even and odd temperature behavior such that the influence of a higher-order dependency is reduced at the cost of an increased influence of a lower-order dependency.

An exchange between two odd or between two even temperature dependencies is also possible. An example of this is depicted in figure 6.3. Curve A in figure 6.3 depicts a reference voltage whose first and second-order temperature dependencies are compensated. When some additional first-order behavior is introduced such that the function rotates, with center T_r, the error band can be reduced, see figure curve B of 6.3. When increasing the first-order dependency, the third-order dependency will also increase slightly. But, as the first-order dependency is much larger than the third-order dependency, a net reduction is obtained.

Thus, the difference between maximum flat and minimum error band is mainly in the reduction of the error by adjusting lower-order dependencies such that the influence of higher-order dependencies is reduced. This gives a reduction in the total error by a factor of about 2 with respect to the method of curve A in figure 6.3. For this, the scaling factors require only a relatively small change; it does not matter for the other design aspects whether a reference is designed according to curve A or curve B in figure 6.3. In the rest of this chapter it is assumed that a reference is made with a maximum-flat output voltage.

6.3.3.2 General set of equations

Assuming a maximum-flat output voltage, the set of equations describing the temperature compensation is given by:

$$\sum_{m=0}^{i} a_m V_{BE_m}(T_r) = V_{REF}, \tag{6.22}$$

$$V_{REF} - \sum_{m=0}^{i} a_m \left[V_G(0)_1 + \frac{kT_r}{q}(\eta - \theta_m) \right] \frac{T - T_r}{T_r} = 0, \tag{6.23}$$

$$\sum_{m=0}^{i} \sum_{n=2}^{\infty} a_m \left[\alpha_n T_r^n - \frac{kT_r}{q}(\eta - \theta_m) \frac{(-1)^n}{n(n-1)} \right] \left(\frac{T - T_r}{T_r} \right)^n = 0. \tag{6.24}$$

In the expression for the first-order compensation, equation (6.24), the constraint for the value of the reference voltage, equation (6.23), is already substituted. To make the possible compensations more clear, some simplifications are made to the equations. As can be seen, the corresponding second and higher-order terms of the m base-emitter voltages are almost equal; the only difference are the θ_ms. When two base-emitter voltages are biased such that their θ_ms are equal, their corresponding second and higher-order terms are equal. Therefore, these higher-order terms are defined as:

$$\text{Higher-order terms of } V_{BE_m} = \sum_{n=2}^{\infty} B_n(\theta_m). \tag{6.25}$$

Assume now, as a start of the discussion of the possible number of orders which can be nullified, that all the $\theta_m = \theta$ are equal, i.e. all the collector bias currents have the same temperature dependency. The constraint for the constant term remains:

$$\sum_{m=0}^{i} a_m V_{BE_m}(T_r) = V_{REF}. \tag{6.26}$$

The sum of the scaled base-emitter voltages at T_r must be equal to the required reference voltage. The constraint for the first-order compensation can be rewritten as:

$$\sum_{m=0}^{i} a_m = \frac{V_{REF}}{V_G(0)_1 + \frac{kT_r}{q}(\eta - \theta)}. \tag{6.27}$$

Thus for a first-order compensation the sum of the scaling factors is given. To have no conflict with equation (6.26), at least two *different* base-emitter voltages are required. Otherwise, the expression for the constant term reduces to:

$$\sum_{m=0}^{i} a_m = \frac{V_{REF}}{V_{BE}(T_r)}, \tag{6.28}$$

which is only fulfilled, including a first-order compensation, when $V_{BE}(T_r) = V_G(0)_1 - \frac{kT_r}{q}(\theta - \eta)$. Of course, this is impractical.

For a higher-order compensation ($n \geq 2$) the following expression must be fulfilled:

$$\sum_{m=0}^{i} a_m B_n(\theta_m) = 0. \tag{6.29}$$

For equal θ_m this reduces to:

$$\sum_{m=0}^{i} a_m = 0. \tag{6.30}$$

This is in contradiction with the first-order compensation. Thus:

> When all the transistors are biased with collector currents having the same temperature behavior, no second or higher-order temperature compensation is possible.

Subsequently, the domain of θ is enlarged to two values. The constraint for the constant term remains the same. The expression for the first-order compensation changes slightly to:

$$V_{REF} - \sum_{m=0}^{i} a_m \left[V_G(0)_1 + \frac{kT_r}{q}(\eta - \theta_m) \right] = 0. \tag{6.31}$$

For the higher-order compensation the $B_n(\theta)$ terms with equal θ can be grouped together. As now two values for θ are possible, the compensation of higher-orders can be split into two summations:

$$\sum_{n=2}^{\infty} \left[\sum_{m=0}^{j} a_m B_n(\theta_0) + \sum_{k=j+1}^{i} a_k B_n(\theta_1) \right] = 0. \tag{6.32}$$

This expression has two solutions. A trivial one:

$$a_m = 0 \quad \forall\, m, \tag{6.33}$$

which contradicts with setting the constant term and the first-order compensation. The other solution is given by:

$$\frac{\sum_{m=0}^{j} a_m}{\sum_{k=j+1}^{i} a_k} = -\frac{B_n(\theta_1)}{B_n(\theta_0)}. \tag{6.34}$$

As the ratio $\frac{B_n(\theta_1)}{B_n(\theta_0)}$ is different for different n, only one higher order can be compensated with two different θs. When it is assumed that the bandgap voltage does not have any second and higher-order temperature behavior, i.e. a

first-order approximation, the $\frac{B_n(\theta_1)}{B_n(\theta_0)}$ terms are all equal for different n and a complete compensation is possible. This is, for instance, assumed in [29]. In this book this approximation is not made. Thus:

> When n different θs are available, n independent polynomials are found for the temperature compensation. Thus the temperature compensation can be done up to the nth order.

Concluding:

- the constant term can always be set;
- for a first-order compensation at least two *different* V_{BE}s are required;
- to be able to compensate n higher-orders (order ≥ 2) at least $n+1$ *different* θs are required.

When it is assumed that for a nth-order compensated bandgap reference the temperature behavior is compensated up to the nth-order the following general conclusion can be made:

> For a n-th-order compensated bandgap reference at least the maximum of (n,2) base-emitter voltages and at least n different θs are required.

As each transistor introduces three parameters that can be used in the compensation:

- $V_{BE_m}(T_r)$;
- θ_m;
- a_m;

more free parameters than minimally required are obtained when the minimum number of transistors is used.

6.3.3.3 First-order compensation

In the previous section the temperature compensation was treated rather generally. In this section the theory is applied to a first-order compensated bandgap reference. As for a first-order compensation at least two base-emitter voltages are required, here a linear combination of two base-emitter voltages is examined. Further, it is assumed that they have equal θs. The reference voltage, V_{REF}, can be written as:

$$V_{REF} = a_1 V_{BE_1} + a_2 V_{BE_2}. \tag{6.35}$$

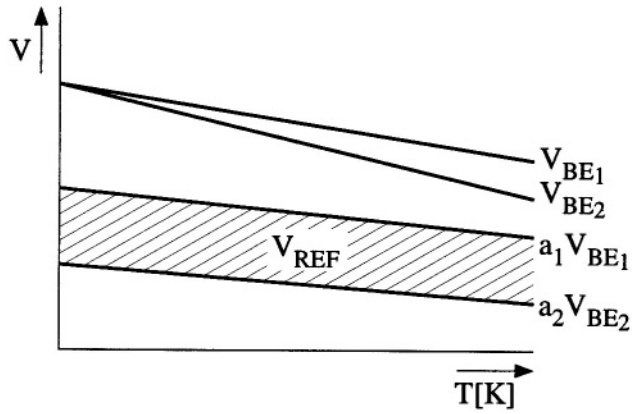

Figure 6.5: The principle of a first-order compensation using two base-emitter voltages.

With equation (6.26) and (6.27) the two scaling factors as a function of the base-emitter voltages can be found:

$$\begin{aligned} a_1 &= \frac{V_{REF}}{V'_G(0)_1} \cdot \frac{V'_G(0)_1 - V_{BE_2}(T_r)}{V_{BE_1} - V_{BE_2}}, \\ a_2 &= -\frac{V_{REF}}{V'_G(0)_1} \cdot \frac{V'_G(0)_1 - V_{BE_1}(T_r)}{V_{BE_1} - V_{BE_2}}, \end{aligned} \tag{6.36}$$

where $V'_G(0)_1 = V_G(0)_1 + \frac{kT_r}{q}(\eta - \theta)$. From these expressions it follows that the two scaling factors have opposite signs. This is necessary for obtaining compensation, as the first-order temperature dependency of a base-emitter voltage is always negative. The principle is depicted in figure 6.5. In the figure the base-emitter voltages are approximated up to their first-order dependency.

After using two degrees of freedom, a_1 and a_2, *two variables still have to be* chosen[1]:

- V_{BE1};
- V_{BE2}.

These can be used for other optimizations, to be discussed later on.

6.3.3.4 Second-order compensation

Again, the bandgap reference with the minimum number of base-emitter voltages is examined. A compensation up to the second order is to be realized

[1] The θs were assumed to be equal and constant, consequently they do not represent degrees of freedom.

and thus at least two base-emitter voltages are required with different θs. The constant term is given by:

$$V_{REF} = a_1 V_{BE_1}(T_r) + a_2 V_{BE_2}(T_r). \tag{6.37}$$

Solving the expressions for a second-order compensation, a ratio for the two scaling factors is found:

$$\frac{a_1}{a_2} = -\frac{B_2(\theta_2)}{B_2(\theta_1)}, \tag{6.38}$$

where θ_1 and θ_2 correspond to V_{BE_1} and V_{BE_2}, respectively. For two given θs the ratio of the two scaling factors is given.

> As ratios depend on matching, the second-order compensation of the bandgap reference depends on the matching instead of on absolute values. Thus a second-order compensation can be implemented relatively accurately.

For the first-order compensation the following constraint is found:

$$V_{REF} - a_1 V'_{G1}(0) - a_2 V'_{G2}(0) = 0, \tag{6.39}$$

in which $V'_{G1}(0)$ and $V'_{G2}(0)$ are the extrapolated bandgap voltages for θ_1 and θ_2, respectively. Combining this expression with the constraint found for the second-order compensation yields for the two scaling factors:

$$a_1 = \frac{V_{REF}}{V'_{G2}(0)} \cdot \frac{1}{\frac{V'_{G1}(0)}{V'_{G2}(0)} - \frac{B_2(\theta_1)}{B_2(\theta_2)}}, \tag{6.40}$$

$$a_2 = -\frac{V_{REF}}{V'_{G1}(0)} \cdot \frac{1}{\frac{B_2(\theta_2)}{B_2(\theta_1)} - \frac{V'_{G2}(0)}{V'_{G1}(0)}}. \tag{6.41}$$

From these two expressions it follows that a_1 and a_2 have, again, different signs and also a previous conclusion is found that two different θs have to be used otherwise the denominator of the last two expressions would become zero.

For the second-order compensated bandgap reference, the two scaling factors are completely determined by the compensation of the first and second-order temperature behavior. For other optimizations to be discussed, these scaling factors can be treated as being constants.

Now the temperature behavior is compensated by choosing the corresponding scaling factors of the base-emitter voltages, the reference voltage still has to be set. Two degrees of freedom are left:

- V_{BE1};
- V_{BE2}.

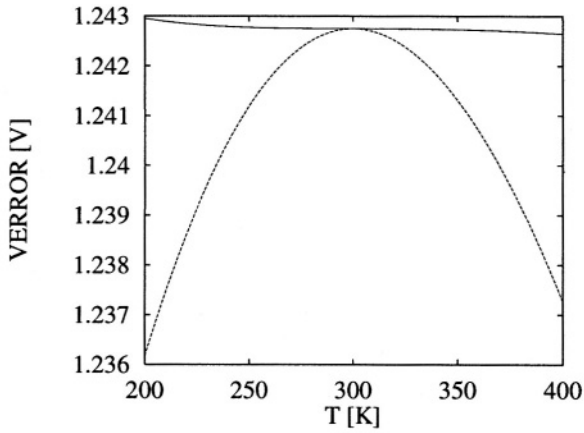

Figure 6.6: The remaining temperature dependency for a first-order (lower function) and a second-order (upper function) compensated bandgap reference with reference voltage equal to $V'_G(0)_1$.

After setting the value of the reference voltage, *one degree of freedom is left* in this equation. Again, this can be used to perform other optimizations.

It should be noted that when a V_{BE} is a degree of freedom in the temperature compensation, it can be equivalent to more than one degree of freedom for other design aspects of the bandgap reference. The base-emitter voltage, for instance, is set by the collector-current, the saturation-current density and the emitter area; in total three parameters. When these three parameters are important for another optimization, three degrees of freedom are obtained for this specific optimization. The only constraint is that the V_{BE} can be freely chosen in the temperature compensation.

The remaining temperature dependencies for a first and second-order compensated bandgap reference are depicted in figure 6.6. For the bandgap voltage as a function of the temperature the expression of Varshni [30] is used:

$$V_G(T) = 1.1557 - \frac{7.021 \cdot 10^{-4} T^2}{T + 1108}. \tag{6.42}$$

The figure clearly shows the mainly second-order temperature dependency of the output voltage of the first-order compensated bandgap reference and the third-order temperature dependency of the second-order compensated bandgap reference. Further, the error voltage of the second-order compensated bandgap reference is considerably smaller than the error voltage of the first-order compensated bandgap reference.

In appendix D a design example can be found of a second-order compensated bandgap reference [31]. The design clearly shows that via the principle

of the linear combination of base-emitter voltages high-performance bandgap references can be obtained.

6.3.4 The key parameters

As may be clear from the previous sections, the design of bandgap references already concerns a lot of parameters when only ideal physical models are used for the base-emitter voltages. For practical bandgap references, the models describing the behavior of the transistor introduce even more parameters. Therefore, it is good to know which parameters of the practical model dominate the behavior of the transistor in the case of bandgap reference design.

The Gummel and Poon model [32] as used in SPICE [33] is a well known model and often used for circuit design. Therefore this model is used here as the basis for the design of bandgap references. A minimum set of key parameters will be derived that describes the relation between the base-emitter voltage and the collector bias current.

The bulk resistors are not taken into account because it is possible to make their influence negligibly small, especially in the case of low current design. The Gummel and Poon model is reduced further to the effects that are relevant for the forward-biased transistor. The leakage currents are ignored too, because in modern IC processes these leakage currents are negligibly small [34]. When these leakage currents cannot be ignored, due to the process characteristics or due to the relatively high temperature at which the bandgap reference has to operate, these leakage currents can be taken into account by using the descriptions as given in [27].

Further, it is assumed that the transistor is biased far from high-level injection (if not, see again [27]). The relevant part of the Gummel and Poon model that remains is given by the following equation [35]:

$$\begin{aligned} I_C(T) &= I_S \cdot \exp\left[\left(\frac{T}{T_{nom}} - 1\right)\frac{E_G}{qV_T}\right] \cdot \left(\frac{T}{T_{nom}}\right)^{X_{TI}} \\ &\times \left[1 - \frac{V_{BC}}{V_{AF}} - \frac{V_{BE}}{V_{AR}}\right] \cdot \exp\left(\frac{V_{BE}}{N_F V_T}\right). \end{aligned} \tag{6.43}$$

The parameters that are used and their meanings are listed in table 6.1.

A further reduction is obtained when the transistor is biased such that $V_{BC} \approx 0$ V. In that case the forward Early effect, modeled by V_{AF}, can be ignored. In contrast to V_{AR} which is on the order of several volts, which cannot be ignored. For a given temperature, V_T is known and T_{nom} is the temperature at which the parameters are extracted from measurements. Finally, N_F equals 1. Thus for an accurate design of bandgap references *four* parameters need to be known accurately, describing the relation between the base-emitter voltage and the collector current (see table 6.2). These parameters are the key parameters for

Parameter	Meaning
V_{BE}	Base-emitter voltage
V_{BC}	Base-collector voltage
I_S	Saturation current
X_{TI}	Order of the temperature dependency of I_S
V_{AF}	Forward Early voltage
V_{AR}	Reverse Early voltage
V_T	Thermal voltage
N_F	Forward-emission coefficient
E_G	Bandgap energy
T_{nom}	Nominal temperature

Table 6.1: The relevant parameters and their meaning in the Gummel and Poon model.

Parameter	Meaning
E_G	Bandgap energy
I_S	Saturation current
X_{TI}	Order of the temperature dependency of I_S
V_{AR}	Reverse Early voltage

Table 6.2: The key parameters for the design of bandgap references.

bandgap reference design. When other models are used, instead of the Gummel and Poon model, the corresponding parameters are found.

6.3.4.1 The bandgap energy

The output voltage of the bandgap reference is directly related to the bandgap energy. An error in the description of the bandgap energy compared to the actual bandgap energy is directly seen in the reference voltage. The method described uses a Taylor expansion of the function describing the bandgap energy. Thus it is largely independent of which function is used for the bandgap energy. In this book the approximation as described by Varshni [30] is used as it is the most commonly used one. It should be noted that in SPICE [33] a first-order approximation of this model is used. However, as can be seen from equation (6.13), both the In-function *and* the bandgap energy are responsible for the second and higher-order temperature behavior of the base-emitter voltage; their contributions appear to be on the same order of magnitude. Therefore, with SPICE-based simulators, second and higher-order compensated bandgap references cannot be simulated accurately.

6.3.4.2 The saturation current and its temperature behavior

The saturation current is one of the parameters determining the base-emitter voltage of a transistor. Via this parameter, spread is introduced in the base-emitter voltage due to emitter-area variations and variations in the doping level or doping profile. In the Gummel and Poon model, X_{TI} models the order of the temperature behavior of the saturation current.

6.3.4.3 The reverse Early voltage

This parameter is the only one modeling a non-ideality with respect to the ideal physical model of the transistor. The reverse Early voltage can be a serious problem. It models the base-width modulation at the base-emitter junction. As the doping level of the base is much lower than the doping level of the emitter, the variation of the depletion layer is predominantly found in the base region, in contrast to the base-width modulation at the base-collector junction where the variation is mainly found in the collector region. The reverse Early voltage can easily be on the order of just a few volts, e.g. 4V, which gives a reduction of the collector current of approximately (see equation (6.43)):

$$\frac{\Delta I_C}{I_C} = 0.15. \tag{6.44}$$

For accurate circuit design, therefore, the reverse Early effect also has to be taken into account. The error due to the reverse Early voltage in the output

voltage of the reference source is derived in [36]. It is given by:

$$\frac{V_{ERROR}}{V_{REF}} = \frac{kT}{qV_{AR}}. \tag{6.45}$$

This error is comparable to errors introduced by the spread in base-emitter voltages due to processing, et cetera [36]. Clearly, in contrast to what the name *reverse* Early voltage suggests, this parameter is also important for the forward mode of the transistor.

6.3.5 Temperature-dependent resistors

Besides the key parameters found in the previous section, one additional phenomenon has to be taken into account. This is the resistance by which the collector current is derived from a voltage. As was discussed in Chapter 3, reference currents are not available in nature and are therefore derived from reference voltages by means of a resistance, for instance. When this resistance is temperature dependent, it introduces extra temperature dependency in the reference current.

Assume a collector bias current, I_C, is derived from a reference voltage, V, by a resistor, R, having a temperature-dependent relative error, as given by:

$$R(T) = R_0[1 + \alpha_1 T' + \alpha_2 T'^2], \tag{6.46}$$

where $T' = T - T_0$, R_0 is the resistance at the nominal temperature T_0 and α_1, α_2 are the first and second-order temperature dependencies of the resistor, respectively. Then, for the collector current the following expression can be found:

$$I_C(T) = \frac{V(T)}{R_0}[1 - \alpha_1 T' + (\alpha_1^2 - \alpha_2)T'^2 + O(T'^3)], \tag{6.47}$$

in which $V(T)$ may be temperature dependent. Then, for the error in the base-emitter voltage the following expression can be found:

$$\begin{aligned} \Delta V_{BE}(T, \alpha_1, \alpha_2) &= V_{BE}(T, \alpha_1, \alpha_2) - V_{BE}(T, \alpha_1 = 0, \alpha_2 = 0) \\ &= \frac{kT_0}{q}[-\alpha_1 T' + (\frac{1}{2}\alpha_1^2 - \alpha_2 - \frac{\alpha_1}{T_0})T'^2 + O(T'^3)] \\ &= \frac{kT_0}{q}[\gamma_1 T' + \gamma_2 T'^2 + O(T'^3)]. \end{aligned} \tag{6.48}$$

Thus a relative error in the resistor causes an additive error in the base-emitter voltage (which is a consequent of the ln-function). The error is independent of the type of collector current intended, i.e. temperature dependent, temperature independent, et cetera, and thus it is the same for each base-emitter voltage.

Resistor type	$\frac{kT_0}{qV_G(0)}\gamma_1[/^oC]$	$\frac{kT_0}{qV_G(0)}\gamma_2[/^oC^2]$
Diffused	$-37 \cdot 10^{-6}$	$-0.21 \cdot 10^{-6}$
Thin-film Nichrome	$-1.1 \cdot 10^{-6}$	$-3.6 \cdot 10^{-9}$

Table 6.3: The resulting relative error at the output of the bandgap reference due to the temperature dependency of the resistors in the V-to-I converter.

Recalling that a bandgap reference is a linear combination of base-emitter voltages, the resulting error at the output of the bandgap reference source can be found. This error voltage, V_{ERROR}, is found from:

$$V_{ERROR} = \sum_{m=1}^{i} a_m \Delta V_{BE_m}(T, \gamma_1, \gamma_2). \tag{6.49}$$

As the error in the base-emitter voltage is equal for all the base-emitter voltages, the relative error equals:

$$\frac{V_{ERROR}}{V_{REF}} = \frac{kT_0}{qV_G(0)_1}[\gamma_1 T' + \gamma_2 T'^2 + O(T'^3)], \tag{6.50}$$

for which equation (6.27) is used with the assumption that the influence of the different θ_ms can be ignored, resulting in $\sum_{m=1}^{i} a_m = \frac{V_{REF}}{V_G(0)_1}$. The final relative error depends on, of course, the type of resistors being used. In table 6.3 examples are given for a diffused resistor and a thin-film Nichrome resistor with respectively, $\alpha_1 = 1.7 \cdot 10^{-3}/^oC$, $\alpha_2 = 5.4 \cdot 10^{-6}/^oC^2$ and $\alpha_1 = 5 \cdot 10^{-5}/^oC$, $\alpha_2 = 0$ (data from [37]). For $\frac{kT_0}{qV_G(0)_1}$ the value 0.026/1.2 is used. The second-order error resulting from the temperature behavior of the diffused resistor is about a factor 4 lower than the second-order behavior of the intrinsic base-emitter voltage. Therefore, when designing second-order (or higher) compensated bandgap references, this effect has to be taken into account. This can be done by adding the corresponding term to the term describing the second-order behavior of the base-emitter voltage.

6.4 Relation to the fundamental design aspects

Ideally, the output voltage of a bandgap reference contains no information, i.e. it is an information-free source. Practical bandgap references do not need to be ideal. Their specifications are determined by the application in which they have to be used and are related to the information processing capacity of that application. Thus the bandgap reference is allowed to have some entropy.

As indicated in Chapter 2, a design should be orthogonalized with respect to noise, bandwidth and signal power. In this section these fundamental design aspects are related to the design of bandgap references.

Noise The choice of the order of the bandgap reference, first or second-order for instance, determines the minimum attainable *systematic* error of the bandgap reference; by using all the key parameters in the design, extra *systematic* errors can be kept to a minimum. The *stochastic* errors are caused by the devices in the bandgap reference, i.e. the transistors and resistors which introduce thermal noise and shot noise. These errors can be kept to a minimum by structured design, i.e. minimization at the mathematical level. Further, process variations also introduce *stochastic* errors in the reference voltage. However, as these errors are time-independent, they can be reduced by using trimming.

Bandwidth In principle, the output power of the bandgap reference is completely located at DC. The remaining part of the spectrum should not contain any power. But, since the bandgap reference is used as a voltage source, its output impedance should be kept at an acceptably low value over a specified bandwidth. Over this bandwidth, the bandgap reference produces noise and therefore power is not only located at DC. The bandwidth of the bandgap reference is determined by the bandwidth that its output impedance must have.

Signal power The signal power of the source at DC is directly given by the specifications. The efficiency of the supply of this signal power is improved when the power-supply voltage is lowered to the reference voltage.

Orthogonalization To meet the required specifications optimally in a relatively short time, the design aspects should be orthogonalized. As was seen in the previous sections, the core of the bandgap reference is a linear combination of base-emitter voltages. These base-emitter voltages set the practical limit on the quality of the bandgap reference. The bandgap reference can be seen as a circuit processing several base-emitter voltages in order to end up with a reference voltage. The base-emitter voltages are the sources of the reference and the scalers and adders are the processing blocks. When the scalers and adders are assumed to be ideal, the maximum practical quality is found. This structure for the bandgap reference is called the idealized bandgap reference in the remaining part of this book. The systematic error is given by the order of temperature compensation whereas the stochastic errors are caused by the noise introduced by the base-emitter voltages and the variation of the process. These last errors can be again reduced by means of a trimming. Since the output of one of the scalers determines the output impedance of the bandgap reference, the bandwidth capabilities of the bandgap reference are still ideal.

The implementation of the scalers and adders also introduce noise. Again, the influence of the process variations can be reduced by trimming. The (electronic) stochastic noise introduced by the scalers and adders should be kept at an acceptable level. This should preferably be orthogonal to the minimization

of the stochastic noise of the idealized bandgap reference in order to obtain the optimum overall noise performance with a minimum level of design complexity.

In Chapter 3 an electronic circuit was treated as a signal path plus its bias circuit. The internal signals of the bandgap reference, i.e. the scaled base-emitter voltages, vary as a function of temperature and in the case of a low-voltage design, the bias circuit may cause signal-dependent systematic errors. Further, as the output impedances of the bias sources are not ideal, errors will penetrate from the power-supply voltage to the output of the bandgap reference.

In the next sections the design with respect to the noise, the bandwidth and the signal power of a bandgap reference is discussed.

6.5 Noise

As the accuracy and temperature independency of bandgap references increase, the mean errors will now become on the order of a few ppm/K over a temperature range of 100 K to 150 K and the noise performance of bandgap references becomes more and more important. For instance: assume a bandgap reference with an output voltage of 200 mV and a mean temperature dependency of 2 ppm/K. The mean uncertainty due to the temperature dependency then equals only $0.4\ \mu V/K$. When the equivalent noise voltage at the output is higher than this value, the noise is the dominant cause of the uncertainty. This example concerns relatively low-frequency noise. For delta-sigma modulators, the relatively high-frequency noise of the bandgap reference is also important. Since the modulators sample at a relatively high rate, the noise is important over a larger bandwidth.

To minimize the total noise at the output of the bandgap reference several methods can be used:

- for a reference voltage for a measurement instrument, the duration of the sampling of the reference voltage can be increased, resulting in low-pass filtering, but inherently slowing down the system;
- perform filtering by means of a capacitor. This requires a relatively large capacitor as the output impedance of a bandgap reference is, in principle, relatively low. When it is too large to be integrated, the capacitor has to be mounted externally to the chip, introducing an additional component and, possibly, also an additional pin;
- the most convenient method is to minimize the noise of the bandgap reference itself. If the specifications are still not met, the former methods can be used too, but now with less severe constraints.

To be able to minimize the noise level of the bandgap reference, all the noise sources in the bandgap reference need to be transformed to the output. For

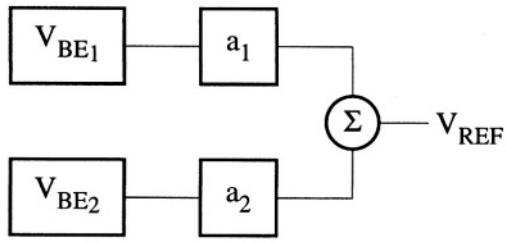

Figure 6.7: A general block diagram of a first and second-order compensated bandgap reference.

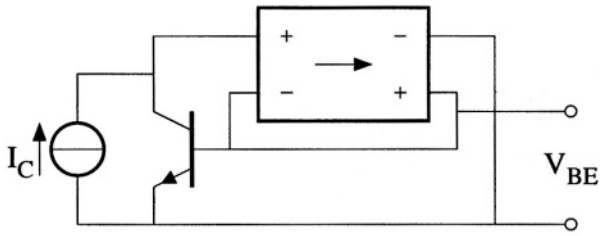

Figure 6.8: An ideal base-emitter voltage generator.

a first and second-order compensated bandgap reference only two base-emitter voltages are required. The general block diagram of those bandgap references can therefore be visualized as depicted in figure 6.7. Three types of blocks can be identified:

- base-emitter voltage generator;
- scaler;
- summing node.

Of these three types of blocks, the base-emitter voltage generators are the core of the bandgap reference. They take care of the required relation to the bandgap energy in order to get a reference voltage. The scalers a_1 and a_2 are dimensionless factors and do not need to introduce any noise. The same goes for the summing node. In contrast, the base-emitter voltages are directly related to the collector current flowing across a junction which is therefore inherently deteriorated by shot noise and consequently the base-emitter voltages are thus always contaminated with noise. In figure 6.8 an ideal base-emitter voltage generator is depicted. The desired collector current, I_C, is forced into the collector by means of negative feedback. The nullor controls the base-emitter voltage such that the desired current flows into the collector. As the input current of

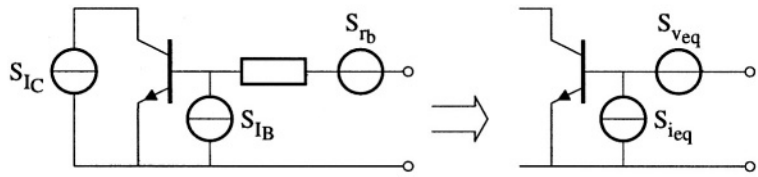

Figure 6.9: The noise sources of a bipolar transistor which have to be transformed to an equivalent noise source at the output of the cell.

the nullor is zero, the complete desired current flows into the collector, and an accurate relation is found between V_{BE} and I_C. Further, as the input voltage of the nullor is zero, it makes it possible to ignore the forward Early voltage. This cell is the core of the idealized bandgap reference and is used to calculate the minimum practical noise level.

6.5.1 Noise of the idealized bandgap reference

To find the noise performance of a single cell, the noise of the transistor has to be transformed to an equivalent source at the output of the base-emitter voltage generator. The transistor noise sources are depicted in figure 6.9. Three noise sources can be distinguished [38]:

- the collector shot noise, $S_{I_C} = 2qI_C$;
- the base shot noise, $S_{I_B} = 2qI_B$;
- the base resistance thermal noise, $S_{r_b} = 4kTr_b$.

For the noise-power density spectrum of the equivalent noise current, $S_{i_{eq}}$, and of the equivalent noise voltage, $S_{v_{eq}}$, see figure 6.9, it holds that [38]:

$$\begin{aligned} S_{i_{eq}} &= 2qI_B\left(1+\frac{1}{\beta_f}\right), \\ S_{v_{eq}} &= 4kT\left\{\frac{1}{2}r_e\left[1+\left(\frac{r_b}{r_e}\right)^2\frac{1}{\beta_f}\right]+r_b\right\}, \end{aligned} \tag{6.51}$$

where β_f is the small-signal forward current-gain factor and r_e equals kT/qI_C. The 1/f noise is ignored in these equations as for modern (bipolar) processes the noise corner can be relatively low. The equivalent noise current does not influence the noise behavior as this source is short-circuited by the nullor output; it is in parallel with a voltage output. Simplifications can be made for the equivalent noise voltage. When the base resistance is made considerably smaller

than r_e, it can be ignored for the noise performance and the equivalent noise source can then be written as:

$$S_{v_{eq}} = 2kTr_e. \tag{6.52}$$

For low-current applications, r_b is very often already much smaller than r_e. For the relatively high-current applications, r_b must be made small by dedicated transistors design, i.e. more and large base-contacts, in order to get optimum performance.

This equivalent noise source can be used for both the base-emitter voltage generators in figure 6.7 and thus for the noise-power density spectrum at the output of the bandgap reference, $S_{v_{out}}$, can be found:

$$S_{v_{out}} = 2kT\left(a_1^2 r_{e1} + a_2^2 r_{e2}\right), \tag{6.53}$$

where a_1, a_2 and r_{e1}, r_{e2} are the corresponding parameters for base-emitter voltages one and two, respectively. This equation describes the noise at the output of the first-order compensated bandgap reference as well as the noise at the output of the second-order compensated bandgap reference. In the following sections this equation will be used to derive the minimum noise productions for the first and second-order compensated bandgap references.

6.5.1.1 Noise of a first-order compensated reference

In Section 6.3.3.3 the scaling factors for a first-order compensated bandgap reference were derived. When these are substituted into expression (6.53), the following expression is found for the noise-power density spectrum of the first-order compensated bandgap reference, $S_{out-first}$:

$$S_{out-first} = 2kT\frac{V_{REF}^2}{V_G'^2(0)} \cdot \frac{[V_G'(0) - V_{BE2}(T_r)]^2 \frac{kT}{qI_{C1}} + [V_G'(0) - V_{BE1}(T_r)]^2 \frac{kT}{qI_{C2}}}{[V_{BE1}(T_r) - V_{BE2}(T_r)]^2}. \tag{6.54}$$

As the noise is inherently caused by the collector currents, the minimization is performed first with the collector currents as the independent parameters. Performing this results in infinite collector currents as in that case the equivalent noise of the base-emitter voltage generators tends to zero. Therefore, a current limitation, I_{MAX}, has to be set:

$$I_{MAX} = I_{C1} + I_{C2}. \tag{6.55}$$

Using this relation between the two collector currents and rewriting the base-emitter voltages as a function of the collector currents and saturation currents, the following expression is found:

$$S_{out-first} = 2q\frac{V_{REF}^2}{V_G'^2(0)} \cdot \left(\frac{T}{T_r}\right)^2 \times$$

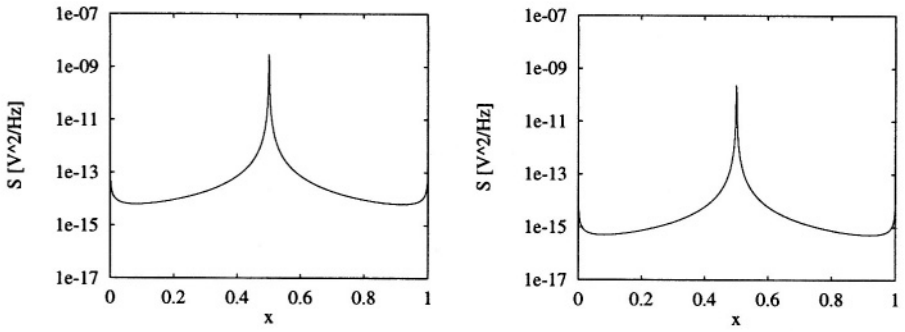

Figure 6.10: The noise power density of a bandgap reference as a function of the collector bias current I_{C1} with $I_{S1} = 16$ aA, $I_{S2}/I_{S1} = 1$ and a) $I_{max} = 1$ μA, b) $I_{max} = 10$ μA $[x = I_{C1}/I_{MAX}]$.

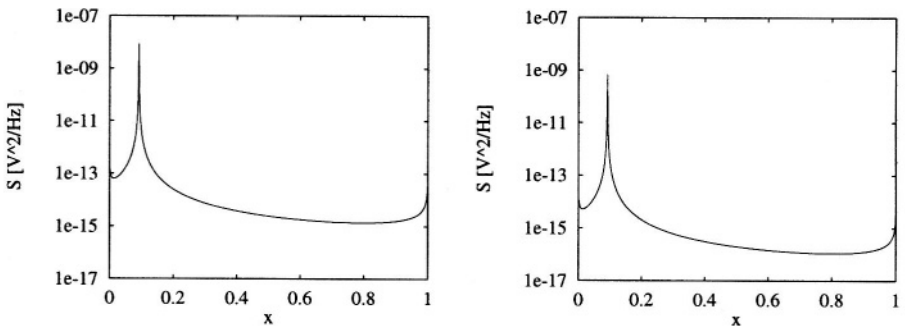

Figure 6.11: The noise power density of a bandgap reference as a function of the collector bias current I_{C1} with $I_{S1} = 16$ aA, $I_{S2}/I_{S1} = 10$ and a) $I_{max} = 1$ μA, b) $I_{max} = 10$ μA $[x = I_{C1}/I_{MAX}]$.

$$\frac{x\left[V_G'(0) - \frac{kT_r}{q}\ln\left(\frac{xI_{MAX}}{I_{S1}}\right)\right]^2 + (1-x)\left[V_G'(0) - \frac{kT_r}{q}\ln\left(\frac{(1-x)I_{MAX}}{I_{S2}}\right)\right]^2}{x(1-x)I_{MAX}\left[\ln\left(\frac{x}{1-x}\cdot\frac{I_{S2}}{I_{S1}}\right)\right]^2}, \quad (6.56)$$

where $x = I_{C1}/I_{MAX}$. From this expression it can be seen that the noise-power density at the output of the bandgap reference is approximately inversely proportional to the total current consumption. This equation is plotted for four situations in figures 6.10a,b and 6.11a,b. In figure 6.10a the saturation currents I_{S1} and I_{S2} are equal and the current consumption is 1 μA. Figure 6.10b shows the noise-power density in the case of equal saturation currents but with a limited current consumption of 10 μA. Figures 6.11a and 6.11b show the noise-power density in the case of scaled emitter areas ($I_{S2}/I_{S1} = 10$) with a

current limitation of 1 μA and 10 μA, respectively. In the four situations, the following values are used for the constants: $V_{REF} = 0.2$ V, $V_G'(0) = 1.19$ V and $T = T_r = 300$ K.

Clearly, the noise-power density has a minimum. The difference between figures 6.10a and 6.10b and between figures 6.11a and 6.11b is the level of the noise-power density; the shape of the function is the same.

According to the figures 6.10a,b and 6.11a,b, the shape of the function is independent of I_{MAX}, and thus the *location* of the minimum is independent of I_{MAX}.

A difference between figures 6.10a,b and 6.11a,b is that the peak in the middle of figures 6.10a,b is shifted out of the middle as in figures 6.11a,b. This is due to the denominator of equation (6.56). The denominator contains a factor:

$$\left[\ln\left(\frac{x}{1-x}\cdot\frac{I_{S2}}{I_{S1}}\right)\right]^2. \tag{6.57}$$

When x equals:

$$x = \frac{I_{S1}}{I_{S1}+I_{S2}}, \tag{6.58}$$

the denominator of equation (6.56) is zero, which results in an infinite noise-power density. Equation (6.58) is fulfilled when the two base-emitter voltages are equal; then the scaling factors a_1 and a_2 become infinite, see equation (6.36), and the noise of the two transistors is infinitely amplified[2]. As the ratio I_{S2}/I_{S1} is different for figures 6.10 and 6.11, a different value for x is found for which the two base-emitter voltages are equal.

The peaks at both borders of the figures are caused by a collector current that is becoming very small. A very small collector current corresponds to a very large equivalent noise-power density at the input of the transistor, see equation (6.52), resulting in a very large noise-power density at the output of the bandgap reference.

Figures 6.10 and 6.11 suggest that an optimum ratio of the two collector currents I_{C1}/I_{C2} can be found and this optimum ratio is independent of I_{MAX}. In [39] an approximated *implicit* equation is derived for the global minimum, i.e. the right local minimum in figures 6.10 and 6.11. The optimum ratio can be found from the following implicit equation:

$$\ln\left(y\frac{I_{S2}}{I_{S1}}\right) = -2\frac{1+y}{1-y}, \tag{6.59}$$

where $y = I_{C1}/I_{C2}$. Thus only the ratio of the two bias currents and the ratio of the two saturation current appear in the expression and, consequently,

[2]This is caused by the fact that with two equal base-emitter voltages it is not possible to realize a first-order compensated bandgap reference with an output voltage not equal to zero.

Figure	I_{S1}/I_{S2}	y_{opt} approximated	y_{opt} exact	$\overline{v_n}[nV/\sqrt{Hz}]$ approximated	$\overline{v_n}[nV/\sqrt{Hz}]$ exact
6.10a	1	10.99	11.36	79.5	79.5
6.10a	1	10.99	11.49	22.5	22.5
6.11b	10	3.55	3.45	39.0	39.0
6.11b	10	3.55	3.44	11.1	11.1

Table 6.4: The optimum ratio y of collector currents for a given ratio of saturation currents and the corresponding minimum noise level for a first-order compensated bandgap reference with $V_{REF} = 200$ mV.

determine the location of the minimum. It also appears to be independent of the reference voltage. Equation (6.59) is solved for the examples in figures 6.10a,b and 6.11a,b. Table 6.4 gives the optimum ratio of the collector currents for a given ratio of the saturation currents, at which the noise-power density is minimal. The noise level, $\overline{v_n}$, and the exact solutions are also given. Thus an optimum ratio I_{C1}/I_{C2} follows from a choice of the ratio of the saturation currents. Further, from the figures it follows that this minimum is relatively flat, especially for larger ratios of the saturation currents. The question which remains is what must be chosen for this ratio: small, large or close to one. This can be derived straightforwardly from equation (6.56), see Appendix E. It appears that the noise power density is a *maximum* for:

$$\frac{I_{S1}}{I_{S2}} = \frac{I_{C1}}{I_{C2}}. \tag{6.60}$$

This is again the condition for equal base-emitter voltages. For practical values of the ratio I_{S1}/I_{S2} it holds that the noise level is steadily decreasing when going away from this maximum, see Appendix E. Thus this ratio has to be as large or as small as possible. However, from the plots of the noise-power density, figures 6.10a,b and 6.11a,b, it follows that when $I_{C1} > I_{C2}$, the condition for the saturation currents has to be $I_{S2} > I_{S1}$. Thus:

- the scaling ratio of the two transistors has to be as large as possible;
- the transistor with the largest current has the smallest emitter area.

The difference of the base-emitter voltages is then the largest for the collector-current ratio and saturation-current ratio which is favorable for low noise [see equation (6.56), the denominator contains a factor $\ln\left(\frac{x}{1-x} \cdot \frac{I_{S2}}{I_{S1}}\right)$ which is proportional to the base-emitter voltage difference].

The same kind of argumentation can be given for the absolute values of the two saturation currents, I_{S1} and I_{S2}. Inspection of equation (6.56) shows that

only terms in the numerator depend on the absolute values of the saturation currents. The terms between the square brackets are equal to $V'_G(0) - V_{BE}(T_r)$. The base-emitter voltage has to be as close as possible to the bandgap voltage, this means that for the practical situation the saturation currents must be as small as possible. It can also be seen in another way that the saturation currents must be as small as possible. For a smaller saturation current and a constant collector current, the base-emitter voltage becomes larger and the equivalent noise voltage of the base-emitter voltage remains constant. Thus the signal-to-noise ratio of the base-emitter voltage increases. This seems to conflict with the constraint that the scaling ratio has to be as large as possible, because with a large scaling ratio, a small and a large transistor are used. The large transistor can be reduced in size by also using a small transistor, and thus inherently reducing the scaling ratio. However, the sensitivity of equation (6.56) is larger for a variation in the *ratio* of the saturation currents than for the *absolute values* of the saturation currents. Assume I_{S2} is the largest of the two saturation currents and is reduced to a lower value. Then the numerator of equation (6.56) reduces somewhat because the term $V'_G(0) - V_{BE2}(T_r)$ reduces somewhat. However, the factor in the denominator decreases faster as here a term $V_{BE1}(T_r) - V_{BE2}(T_r)$ decreases more as $V_{BE1}(T_r) < V'_G(0)$. Moreover, the term in the numerator is not a factor of the numerator. Thus, to get optimum noise performance, the smallest and largest allowed transistors must be used.

For the optimum noise performance of first-order compensated bandgap references, the following rules have been found:

- the ratio of the two collector currents, $y = \frac{I_{C1}}{I_{C2}}$, follows from: $\ln\left(y\frac{I_{S2}}{I_{S1}}\right) = -2\frac{1+y}{1-y}$;
- the noise level is inversely proportional to $I_{C1} + I_{C2}$;
- the ratio I_{S1}/I_{S2} should be as large as possible for $I_{C1}/I_{C2} < 1$ and vice versa;
- $I_{S1} + I_{S2}$ should be as small as possible.

In appendix F a design example of a first-order compensated bandgap reference [39] can be found. This example shows that the noise of a realistic bandgap reference can be considerably close to the noise of the idealized bandgap reference.

6.5.1.2 Noise of a second-order compensated reference

For the noise performance of second-order compensated bandgap references the corresponding scaling factors, equations (6.40) and (6.41), have to be used in equation (6.53). But, as these scaling factors are already completely determined by the first and second-order temperature compensation, they are now constants

for the noise minimization. Therefore, the equation for the noise minimization is given by:

$$S_{out-second} = 2qV_T^2 \left(\frac{a_1^2}{I_{C1}} + \frac{a_2^2}{I_{MAX} - I_{C1}} \right). \tag{6.61}$$

The noise-power density is minimal for the following optimum ratio of collector currents:

$$\frac{I_{C1}}{I_{C2}} = -\frac{a_1}{a_2} = \frac{B_2(\theta_2)}{B_2(\theta_1)}. \tag{6.62}$$

The ratio $+\frac{a_1}{a_2}$ is also a solution, but as either a_1 or a_2 is negative, the resulting ratio is negative and this is not practical. The two corresponding collector currents equal:

$$\begin{aligned} I_{C1} &= \frac{a_1}{a_1 - a_2} I_{MAX}, \\ I_{C2} &= \frac{-a_2}{a_1 - a_2} I_{MAX}. \end{aligned} \tag{6.63}$$

Substitution of the expressions for a_1 and a_2, equations (6.40) and (6.41), the following expression is found:

$$\begin{aligned} S_{out-second} &= 2qV_T^2 \left[\frac{V_{REF}}{V_G'(0)} \right]^2 \cdot \left[\frac{B_2(\theta_1) + B_2(\theta_2)}{B_2(\theta_1) - B_2(\theta_2)} \right]^2 \cdot \frac{1}{I_{MAX}} \\ &= V_{REF}^2 \cdot \frac{CONSTANT}{I_{MAX}} \end{aligned} \tag{6.64}$$

in which it is assumed that $V_{G1}'(0) = V_{G2}'(0) = V_G'(0)$, resulting in a negligibly small error. From this expression some remarkable conclusions can be drawn:

- the noise-power density of a second-order compensated bandgap reference with a given reference voltage, can only be influenced by the designer by means of the current consumption; it is inversely proportional to the total current consumption;
- for a given current consumption the signal-to-noise ratio is the same for each reference voltage;
- the size of the transistors used does not influence the noise level as it did for the first-order compensated bandgap references.

The noise voltage is on the order of:

$$\overline{v_{eq}} = V_{REF} \cdot 10^{-10} \sqrt{\frac{B}{I_{MAX}}}, \tag{6.65}$$

in which it is assumed that the noise spectrum is white. From this expression the minimum current consumption can easily be found for a given reference voltage and a required noise level.

Example: Assume a second-order compensated bandgap reference with an output voltage equal to 1 V is required with a noise voltage that is at most $1\ \mu V/\sqrt{Hz}$. From expression (6.65) follows a minimum current consumption of 10 nA.

6.5.2 The noise of the remaining parts

In the previous discussion only the base-emitter voltage generators were assumed to introduce noise. When it is not possible to reach the required performance at that level of hierarchy, the bandgap reference will not reach its performance without the use of filters since in the subsequent design levels, in which other blocks are implemented, the performance only degrades. These other blocks are:

- the adder;
- the scaler.

These blocks will introduce noise when they are implemented by realistic circuits. Further, when the design of the signal part of the bandgap reference is finished (i.e. the base-emitter voltage generators, scalers and adders), the biasing of the complete bandgap reference has to be done [40]. As the bias sources in a low-voltage environment cannot be made relatively ideal, the noise of the bandgap reference will increase and, due to the finite output impedance of the bias-current sources, variations from the power-supply voltage penetrate to the output of the bandgap reference. Therefore, the following are also discussed:

- biasing;
- power-supply rejection.

The design with respect to the adders, scalers and biasing circuit should preferably be orthogonal to the design of the idealized bandgap reference as the minimization of its noise level then remains valid.

6.5.2.1 Noise of an adder

The choice of whether to use an active adder or a passive adder is important for the output impedance of the bandgap reference and the interactions of the blocks of the bandgap reference. The passive and active addition of two voltages is depicted in figure 6.12. In this figure a negative voltage is added to a positive one, as this is a common situation for bandgap references. In both cases the negative voltage source is floating. This means that the voltage source is only connected via current sources to the supply rails. When mismatches occur between the currents of these sources, i.e. the inward current is not equal to

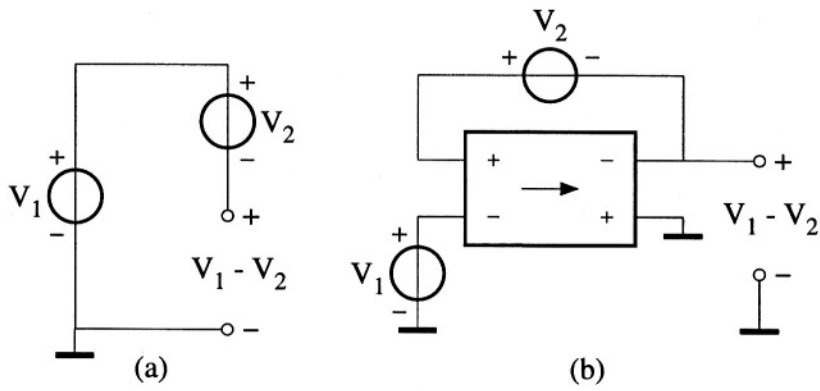

Figure 6.12: a) Passive and b) active addition of two voltages.

the outward current, the resulting offset current will flow out of the voltage-source terminals into the other voltage source. When the output impedance is not relatively low, a relatively large error voltage is introduced. For the active addition this offset current is supplied by the nullor and, consequently, the error voltage is zero.

In the case of the passive addition, the noise is mainly a stochastic DC offset voltage due to the mismatch errors in the current sources. Also, when the currents are noisy, additional noise is found due to the shot noise of the current sources.

For the active addition the noise introduced is from the input stage of the nullor implementation. For a (MOS)FET input stage, the equivalent noise power is inversely proportional to the bias current. For a bipolar input stage the noise can also be said to be inversely proportional to the bias current. However, a minimum can be found which is located at a relatively large bias current (see Section 3.4) as the "source" impedance experienced by the transistor is relatively low.

Thus, for a passive adder the noise is increased due to offset currents and their possible shot noise. For the active adder the noise is also increased. The extra noise power is more or less inversely proportional to the current consumption of the input stage. As the noise contribution of the base-emitter voltage generators is also inversely proportional to the current consumption, an optimal division of the current may be found independent of the previous noise consideration to end up with an overall minimum noise level.

6.5.2.2 Noise of the scalers

When the bandgap reference is completely realized in the analog domain, the sealers are, very likely, implemented by means of resistive ratios thereby in-

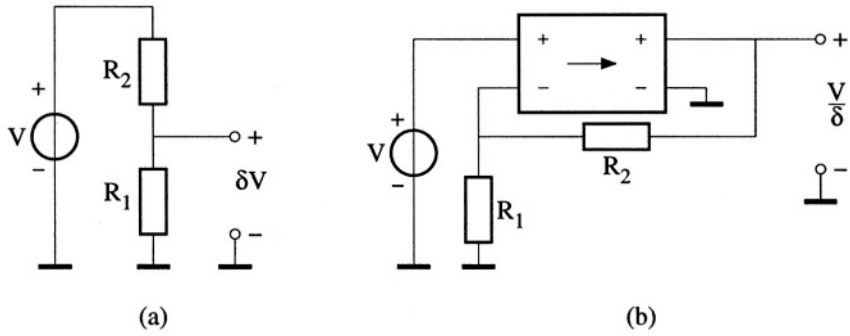

Figure 6.13: Two voltage scalers, a) $a < 1$ and b) $a > 1$.

troducing thermal noise. When, however, only the base-emitter voltages are integrated and the scaling and addition is done in the digital domain [24], the resolution of the digital words introduces errors.

Here it is assumed that the sealers are realized in the analog domain. Ideally, DC transformers would do the job but these do not exist. Theoretically, capacitive dividers with the correct initial conditions could also do the job. However, due to leakage currents the capacitor voltage will steadily increase/decrease as a function of time and will finally end up clipped to a supply rail. To guarantee zero currents is, of course, impossible as the bandgap reference is intended for use in a wide range of temperatures and, consequently, the leakage currents are always present. Therefore, the sealers are most likely to be realized by resistances.

The noise behavior of the scalers depends on the type of bandgap reference, first-order compensated or second-order compensated. For both types it is assumed that the scaling factors are realized by resistive ratios. In figure 6.13, the two possible sealers are depicted, $a < 1$ and $a > 1$. Both scalers have a resistive divider. In figure 6.13a, the divider is in the forward path and thus the transfer is smaller than one. In figure 6.13b, the resistive divider is in a feedback path and the transfer is therefore larger than one.

A bandgap reference comprises at most two of these voltage scalers. In order to find a non-infinite current for the minimum noise level (cf. the noise-minimization of the base-emitter voltage generators), a maximum current consumption is set, I_{MAX}. This current is the sum of the currents flowing through the two resistive dividers in the bandgap reference. Transforming the resulting noise to the output of the bandgap reference, the noise level can easily be minimized. The noise-power density at the output of the bandgap reference

(first-order and second-order compensated) is given by:

$$S_v = \frac{4kT}{I_{MAX}} \cdot \left(\frac{a_1|1 - a_1|V_{BE1}}{\gamma} + \frac{-a_2|1 + a_2|V_{BE2}}{1 - \gamma} \right), \qquad (6.66)$$

where γ is the ratio of I_{C1} and I_{MAX}. In this equation a_1 is assumed to be positive and a_2 to be negative. The modulus signs make the equation valid for all scaling factors, i.e. larger, equal and smaller than one. From this expression can be seen that the noise of the resistive dividers at the output of the bandgap reference is inversely proportional to I_{MAX}. The larger this current is, the lower the impedance level and the lower the noise-power density. Parameter γ can be used to minimize the noise contribution.

For the first-order compensated bandgap reference, the noise minimization of the idealized bandgap reference has an influence on the scaling factors. Thus γ following from the minimization of equation (6.66), with a_1 and a_2 assumed to be constant, does not need to result in the global minimum.

For reference voltages below approximately 0.5 V, by choosing the appropriate base-emitter voltages it is possible to find scaling factors close to one. In that case the noise due to the scaling factors is considerably reduced. From the minimization of the noise of the idealized bandgap reference other values may be found for the scaling factors, that are not as close to one as would be possible. However, for a ratio of the saturation currents beyond about ten, the minimum of the idealized bandgap reference noise is relatively flat, see figure 6.11. In that case, some freedom is obtained in the choice of the scaling factors without having much influence on the noise level of the idealized bandgap reference. Then both are near their minimum. The global minimum will in that case not be far from this noise level.

The strategy for reference voltages below about 0.5 V should therefore be to find a convenient scaling ratio for the saturation currents for which the idealized bandgap reference has a relatively flat minimum noise level and realize scaling factors which are close to one.

For higher reference voltages, it is best to keep the difference of the base-emitter voltage as large as possible since the scaling factors then become the smallest (closest to one seen from the side larger than one) resulting in the lowest noise level.

For the second-order compensated bandgap references the situation is a bit different. As the two scaling factors for these bandgap references are completely determined by the first and second-order temperature compensation, they can be assumed to be constant for the minimization of the scaling factor noise. The minimization of the noise with respect to γ is the only thing that can be done. The choice of the base-emitter voltage has only a slight influence on the noise level of the scaling factors. The base-emitter voltages only vary by a few hundred milli-volts. Further, when V_{BE1} increases, V_{BE2} reduces in order to keep the

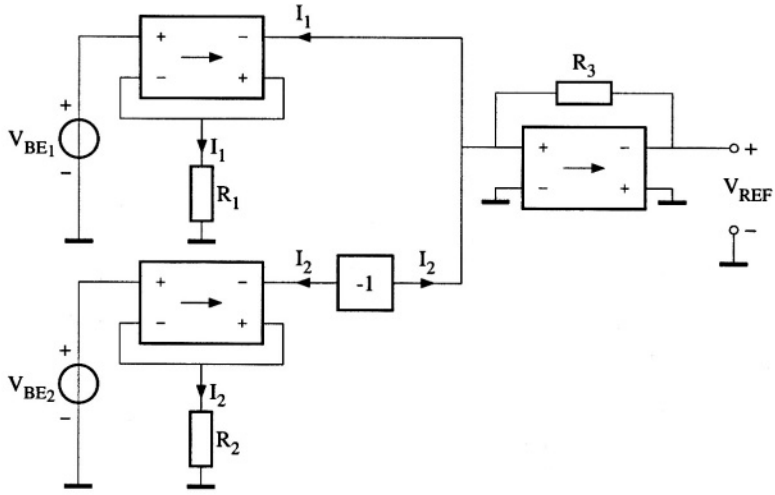

Figure 6.14: A bandgap reference with the scaling realized in the current domain.

correct reference voltage, see equation (6.37). Performing the minimization of equation (6.66) with respect to γ, the bandgap reference noise will be close to the global minimum. A convenient expression for γ_{opt} can be found with the help of the following substitution:

$$\gamma = \frac{x}{1+x}, \tag{6.67}$$

where x is the ratio of I_1 and I_2 with the constraint that $I_1 + I_2 = I_{MAX}$. Performing a minimization with respect to x yields:

$$x_{opt} = \sqrt{\frac{a_1|1 - a_1|V_{BE1}}{-a_2|1 + a_2|V_{BE2}}}, \tag{6.68}$$

Subsequently, γ_{opt} is found from substituting x_{opt} in equation (6.67).

Of course, the division of the current between the idealized bandgap reference and the scaling factors still has to be done; both are inversely proportional to the current consumption. This is a straightforward minimization.

Up to now, the scalers were assumed to be dimensionless. However, when the reference voltage becomes relatively close to the power-supply voltage, it may be necessary to do the scaling and addition in the current domain to be able to realize the required reference voltage. In figure 6.14 a bandgap reference is depicted which uses two transconductance amplifiers and one transimpedance amplifier to perform the scaling in the current domain. The inverter is required

to obtain one negative scaling factor. As the scaling factors that need to be realized in order to achieve the required temperature compensation and reference voltage are independent of the domain in which they are realized, the idealized bandgap reference noise is the same for this structure and the one with scaling in the voltage domain.

A difference can be found for the scalers. Assume the current through resistor R_1 is I_1 and the current through resistor R_2 is I_2. Then the current through R_3 equals $I_1 - I_2$. The total required current can be found as follows. The inverter supplies I_2 at both its input and output. The output of the transimpedance amplifier supplies a current equal to $I_1 - I_2$. This current flows into the summing node, together with I_2 from the output of the inverter, resulting in an out-flowing current equal to I_1 which flows into the output of the first transconductance amplifier. Thus the total current consumption for the scalers is :

$$I_{MAX} = I_2 + I_2 + I_1 - I_2 = I_1 + I_2. \tag{6.69}$$

This total current consumption required by the scalers is the same as it is the bandgap reference scalers implemented in the voltage domain.

For the noise at the output of the reference caused by the resistors of the scalers, S_v, the following expression can be found:

$$S_v = 4kT\left(\frac{R_3^2}{R_1} + \frac{R_3^2}{R_2} + R_3\right). \tag{6.70}$$

For the two scaling factors the following hold:

$$a_1 = \frac{R_3}{R_1}, \tag{6.71}$$

$$a_2 = -\frac{R_3}{R_2}, \tag{6.72}$$

and for R_1 and R_2 the following expressions hold:

$$R_1 = \frac{V_{BE1}}{\gamma I_{MAX}}, \tag{6.73}$$

$$R_2 = \frac{V_{BE2}}{(1-\gamma) I_{MAX}}, \tag{6.74}$$

where γ is again the ratio between I_1 and I_{MAX}. Using these four relations, the expression for the noise-power density at the output of the bandgap reference can be rewritten as:

$$S_i = \frac{4kT}{I_{MAX}}\left(\frac{a_1^2 V_{BE1}}{\gamma} + \frac{a_2^2 V_{BE2}}{1-\gamma} + \frac{V_{REF}}{2\gamma - 1}\right). \tag{6.75}$$

When comparing this equation with equation (6.66), it is easily seen that this noise level is always higher than the one described by equation (6.66). Further, the impedance levels of the two scalers can no longer be chosen independently as they are coupled via resistor R_3. The noise minimization with respect to γ is considerably reduced.

6.5.2.3 Biasing

The final bandgap reference requires bias currents for proper functioning. At the least, it requires two current sources for biasing the two of the base-emitter voltage generator transistors. As shown in [35], a bias current source, without series feedback, has a noise-power density at its output, S_i, of about:

$$S_i = 2qI_C, \tag{6.76}$$

where I_C is the bias current and q is the Boltzmann constant. In the case of current mirrors used for biasing, this value is doubled (for a mirror factor equal to one). This noise is directly added to the collector current of the reference transistors. Therefore, the noise-power density of the reference transistors doubles to:

$$S_v = 4kTr_e. \tag{6.77}$$

As the noise-power density of both transistors doubles, the noise-power density of the complete idealized bandgap reference is doubled. The noise minimum of the idealized bandgap reference is not influenced by this extra noise as it can be accounted for by an overall factor. Of course, the noise of the resistive part of the voltage scalers remains the same. To obtain the original noise level again, three possibilities exist:

- using series feedback in the current sources;
- increasing the bias currents;
- using a voltage multiplier for the power supply.

Using series feedback in current sources When it is possible to use series feedback in the current source, the noise can be reduced. For a voltage across the series-feedback resistor of n times kT/q, the thermal voltage, the noise at the output can be approximated by [41], [35]:

$$S_i = \frac{2qI_C}{(1+n)^2} \quad \vee n < \sqrt{\beta}, \tag{6.78}$$

where β is the current-gain factor of the transistor in the current source. For n=2, a voltage of about 50 mV is required across the series resistors in the

current source and the equivalent noise-power density at the output of the base-emitter voltage generator equals:

$$S_v = 2kTr_e\left(1+\frac{1}{(1+n)^2}\right) \approx 2.2kTr_e. \tag{6.79}$$

The noise-power density is increased by only ten percent. For a second-order compensated bandgap reference with a supply voltage of 1 V, intended for a temperature range on the order of 100 K, this 50 mV is very easily too large a reduction. Each reduction of the effective supply voltage for the active part of about 2 mV means a reduction by 1 K in the functional temperature range; 50 mV extra for the current sources means 25 K less for the functional temperature range of the bandgap reference.

Increasing the bias currents When the bias current of the reference transistor is doubled, the equivalent noise-power density at the output of the base-emitter voltage generator is halved, see equation (6.77). The original noise level is obtained again at the cost of double the power consumption.

Using a voltage multiplier for the power supply As was already discussed, when it is possible to use a series feedback in the current source, its output noise can be reduced considerably. With a voltage multiplier, the power-supply voltage can be increased by 100 mV, for instance, making the noise of the current source negligibly when this 100 mV is used for series feedback. Of course, only the nodes of the current sources that feed the reference transistors need to have a higher supply voltage. When the efficiency of the voltage multiplier is about 50 percent, which is an optimistic guess (see Chapter 3), the power consumption for the reference transistors is increased by a factor of 2.2 (for a 1 V supply). For larger multiplication factors the power consumption increases even further. In this value the power consumption of the auxiliary circuitry required for the voltage multiplier, an oscillator for instance, has not been taken into account.

Thus, voltage multipliers are not an efficient choice when they are used to reduce the noise production of the current sources. They are likely to consume more power than the extra power that would be consumed when the bias current of the reference transistors is increased.

6.5.2.4 Power-supply rejection

Practical current sources will have a finite output impedance. When the supply voltage is not noise free, disturbances from this supply voltage will penetrate via these finite output impedances to the output of the bandgap reference, resulting in extra noise.

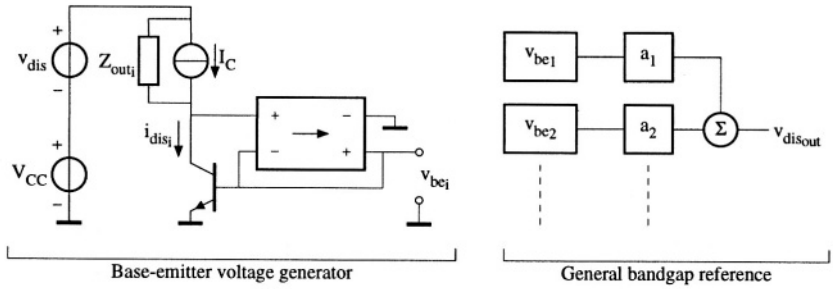

Figure 6.15: An idealized bandgap reference with non-ideal current sources.

Assume the bandgap as given in figure 6.15. The nullors make the collector nodes low-impedance nodes, in order to make the bias currents completely flow into the collector leads. Therefore, for a disturbance, v_{dis}, on the supply-voltage rail, the currents injected into the collector leads of the reference transistors, i_{dis_i}, equal:

$$i_{dis_i} = \frac{v_{dis}}{Z_{out_i}}, \tag{6.80}$$

where Z_{out_i} is the output impedance of the corresponding current source. The low-frequency output impedance of the current sources, r_{out_i}, is given by:

$$r_{out_i} = \frac{V_{AF_i}}{I_{C_i}}. \tag{6.81}$$

It may be assumed that the forward Early voltages are equal for the current sources. The resulting disturbance on the base-emitter voltage is related to i_{dis_i} via $1/g_m$. Then, the disturbances found at the output of the bandgap reference amount to:

$$v_{dis_{out}} = \sum_{i=1}^{2} a_i \frac{v_{dis}}{g_{m_i} \frac{V_{AF}}{I_{C_i}}} = v_{dis} \frac{V_T}{V_{AF}} \sum_{i=1}^{2} a_i = v_{dis} \frac{V_T}{V_{AF}} \frac{V_{REF}}{V'_G(0)}, \tag{6.82}$$

where V_T is the thermal voltage. A commonly used figure of merit is the power-supply-rejection ratio (PSRR); it is a measure for how good the isolation is between the power-supply voltage and the output of a circuit. For the bandgap reference this PSRR is given by:

$$PSRR = 20 \log \left[\frac{V_T}{V_{AF}} \frac{V_{REF}}{V'_G(0)} \right]. \tag{6.83}$$

It should be noted that for a bandgap reference realized with more base-emitter voltages this same expression is found.

Example: $V_T = 25$ mV, $V_{AF} = 50$ V, $V_{REF} = 0.2$ V and $V'_G(0) = 1.2$ V. The PSRR of the bandgap reference is then -82 dB.

For this derivation of the PSRR it was assumed that the current sources are equal; in this case they did not have any series feedback. When series feedback is applied, the PSRR improves as the output impedances of the current sources increase. When it is possible to realize the output impedances such that the injected disturbances of the current sources cancel at the output, a very high PSRR can be achieved. Of course, the ratio of the two output impedances becomes very important and this may be too tough a job to reach the desired matching.

The required measures, however, can be taken independently of the other design consideration as discussed in the previous section. Therefore, optimization with respect to PSRR, at least the low-frequency behavior, is orthogonal to the other design steps.

6.6 Bandwidth

The output signal of the bandgap reference is, ideally, located at DC. This does not mean that the behavior of the bandgap reference at higher frequencies is not important. The bandgap reference is a part of an information processing system which is limited by Shannon's maximum channel capacity. Therefore, the bandgap reference has to perform correctly up to the maximum frequency of interest, i.e. the frequency to which the bandgap reference can have an influence on the system performance.

In the previous sections the noise production of the bandgap reference was minimized. It was assumed that the noise-power density is frequency independent and thus the noise level was minimized over the complete frequency band. When the resulting noise level is too high, it was shown that at the cost of an increased current consumption the noise level can be lowered. However, when an increased current consumption is not allowed, one has to rely on filtering. But, as the noise is minimal for the given current consumption, the additional filtering required is also minimized.

6.6.1 Output impedance

The output impedance of the bandgap reference has to be relatively low otherwise load currents introduce too large variations on the reference voltage. In figure 6.12 two methods were depicted for the addition of two scaled base-emitter voltages. When it is done passively, i.e. the voltage sources are more or less stacked upon each other, the output impedance of the bandgap reference is given by the sum of the output impedances of the two scaled base-emitter

voltage generators. In that case those blocks have to be optimized, not only for the DC behavior, but also for the high-frequency output impedance.

For the active addition, these aspects are orthogonalized. The output impedance of the bandgap reference is realized by the nullor performing the addition. The base-emitter voltage generators can then be optimized for the DC behavior while the adder can realize a wide-band low output impedance.

6.6.2 Power-supply rejection

The expression for the PSRR which was found in the previous section only holds, of course, for the relatively low frequencies. Parasitic capacitances in parallel with the output of a current source will cause an increase in the injected disturbances for higher frequencies. By appropriate scaling of the parallel capacitances some canceling may be obtained but, again, this poses severe constraints on the matching of the capacitances in order to obtain a relatively high PSRR.

To reduce the PSRR for relatively high frequencies, the scaling factors can be made frequency dependent in order to obtain canceling of noise injected from the supply rails. This is, of course, not orthogonal to the design of the noise generation for higher frequencies of the bandgap reference itself. Depending on the realization of the scaling factor, its gain factor can easily be made one or zero by means of a capacitor and this can be favorably used to realize an increased PSRR for higher frequencies.

6.7 Signal power

In order to maximize the dynamic range of the bandgap reference output voltage, the reference voltage must be as large as possible for a given noise level. When discussing the noise behavior of bandgap references, it was found that the dynamic range of a bandgap reference is given when the current consumption is given. Increasing the current consumption is thus the only way to increase the dynamic range. Further, for a given *current* consumption the *power* consumption reduces for lower supply voltages. Therefore, to be as power efficient as possible, the power-supply voltage must be as close as possible to the reference voltage. This implicitly means that the required voltage for the current sources should be as low as possible.

For low-voltage design, i.e. supply voltages on the order of 1 V, the available voltage for the current sources is already very low such that they are close to saturation. Saturation of the current sources leads to a temperature dependent error in the reference voltage. Therefore the internal node voltages of the bandgap reference and the saturation voltages of the transistors in the current sources should be minimized. The influence of the saturating current sources is then minimized for a given power-supply voltage.

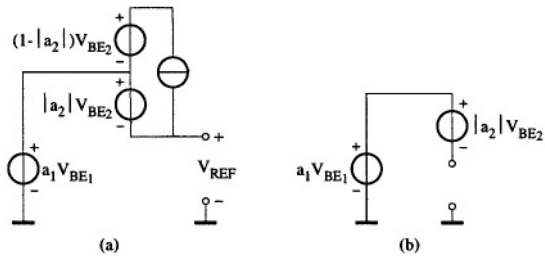

Figure 6.16: Two basic structures for bandgap references in the voltage domain. a) Structure with $|a_2| \leq 1$ and b) with $|a_2| > 1$.

6.7.1 Internal node voltages

Three basic structures can be found for bandgap references realized in the voltage domain. These are:

- $a_1 < 1$ and $|a_2| < 1$;
- $a_1 \geq 1$ and $|a_2| \leq 1$;
- $a_1 \geq 1$ and $|a_2| > 1$;

where a_2 is the scaler with the negative transfer. The case of $a_1 < 1$ with $|a_2| \geq 1$ conflicts with the assumption that a_2 is the negative scaler. For the first structure, V_{BE1} is the highest voltage. Minimizing this voltage is realized by choosing large transistor, as is discussed for the other two structures. The other two structures are depicted in figure 6.16. For the structure of figure 6.16a the highest voltage in the bandgap reference is found at the top of the V_{BE2} generator:

$$V_{MAX} = V_{REF} + V_{BE2}. \tag{6.84}$$

To minimize this voltage, V_{BE2} has to be as small as possible. It should be noted that V_{BE2} is always the smallest of the two base-emitter voltages because a_2 is chosen to be the negative scaler. Reducing the bias current is not allowed as that results in an increased noise level. Thus the size of the transistor has to be chosen as large as possible. For the noise minimization of a first-order compensated bandgap reference this was also found; the noise level of a second-order compensated bandgap reference is not influenced. The other saturation current, I_{S1}, is however related to I_{S2} as a result of constraining a certain value for the reference voltage, according to:

$$I_{S2} = I_{C2} \cdot \left(\frac{I_{S1}}{I_{C1}}\right)^{\frac{B_2(\theta_2)}{B_2(\theta_1)}} \exp\left[\frac{V'_G(0)}{V_T} \cdot \frac{B_2(\theta_2) - B_2(\theta_1)}{B_2(\theta_1)}\right]. \tag{6.85}$$

I_{max} / I_{S1}	$1nA$	$100nA$	$10\mu A$	$1mA$
$1aA$	1881	533	151	43
$10aA$	3535	1001	283	80
$100aA$	6643	1881	533	151

Table 6.5: The ratio I_{S2}/I_{S1} as a function of the current level and I_{S1}.

This may limit the practical usefulness of reducing the minimum power-supply voltage required for a second-order compensated bandgap reference by using large transistors. In table 6.5 the value of I_{S2} is given as a function of the current level and I_{S1}.

For the structure of figure 6.16b the maximum node voltage is found at the summing node. For the first-order compensated bandgap reference this equals:

$$V_{MAX} = a_1 V_{BE1} = \frac{V_{REF}}{V'_G(0)} \left(\frac{V'_G(0) - V_{BE2}}{V_{BE1} - V_{BE2}} \right) V_{BE1}, \tag{6.86}$$

and for the second-order compensated bandgap reference:

$$V_{MAX} = a_1 V_{BE1} = \frac{V_{REF}}{V'_{G2}(0)} \frac{V_{BE1}}{\frac{V'_{G1}(0)}{V'_{G2}(0)} - \frac{B_2(\theta_1)}{B_2(\theta_2)}}. \tag{6.87}$$

For the first-order compensated bandgap reference, the most effective solution is to increase the difference between the two base-emitter voltages. This is in accordance with rules found for noise minimization. For the second-order compensated bandgap reference, a_1 is completely determined by the compensation of the temperature behavior. Thus the only thing that remains is to reduce V_{BE1} by increasing the area of the transistor.

Summarizing, it may be clear that reducing the maximum internal node voltage of a bandgap reference, by means of reducing a base-emitter voltage, is only minimally effective as a base-emitter voltage can only be reduced slightly by altering the saturation currents. Using scaling in the current domain, gives the lowest internal node voltages, i.e. only V_{BE1}, V_{BE2} and V_{REF} are found, at the cost of an increased noise level.

6.7.2 Saturating bipolar transistors

When the internal node voltages are not low enough, the current sources are saturating for the lower temperatures. As this saturation has a very non-linear effect on the current of the corresponding source, the base-emitter voltage shall have an extra non-linear voltage and the temperature independency of the output voltage of the bandgap reference is deteriorated.

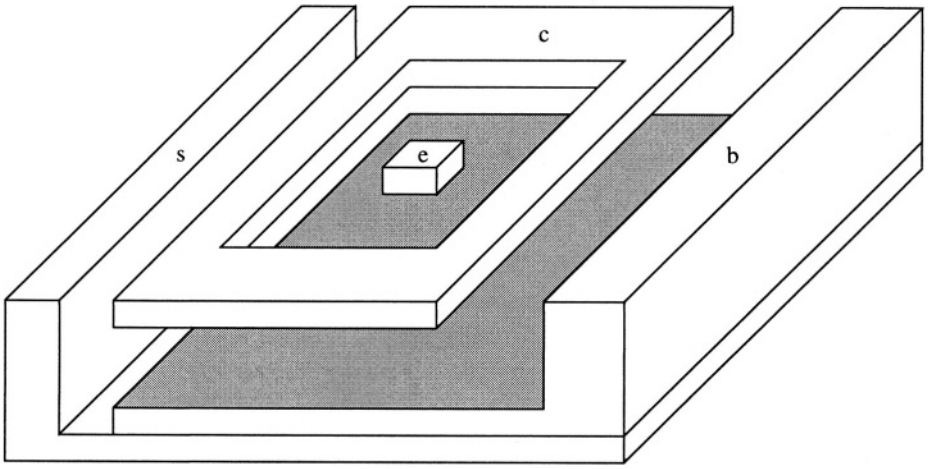

Figure 6.17: A typical structure of a lateral PNP transistor.

In reference [42] the collector-emitter voltage is calculated as a function of the level of saturation. However, this is done for a transistor without a substrate and further, for this purpose the interest is in the collector current instead of in the collector-emitter voltage. Therefore the saturation behavior of bipolar transistors is studied here from this other point of view. As the saturation behavior for the vertical and lateral transistor cannot be described by one model, these are treated separately.

6.7.2.1 The lateral transistor

In a lot of processes the PNP transistors are of the lateral type [43], [44]; a typical structure is depicted in figure 6.17. In an epitaxial area an emitter doping is realized which is completely surrounded by a collector doping in order to collect, in the normal forward mode, the most of the laterally injected carriers from the emitter; the base is connected via a low-ohmic buried layer. In the normal forward mode, the base region is injecting carriers into the emitter, which is one of the causes of the finite current-gain factor of the transistor. In the reverse mode, the collector injects carriers into the base, and the emitter then has to collect them. But as the emitter does not surround the collector, a lot of carriers flow to the substrate, especially the carriers injected at the outside of the collector.

For a transistor in the normal forward mode, this reverse mode transistor starts conducting when the transistor starts saturating, resulting in a change of the collector current. For a transistor without a substrate, the collector current,

I_C, can be written as:

$$I_C = I_S \left[\exp\left(\frac{V_{EB}}{V_T}\right) - \exp\left(\frac{V_{CB}}{V_T}\right) \right] \cdot \left[1 + \frac{V_{BC}}{V_{AF}} \right], \qquad (6.88)$$

where I_S is the saturation current of the *junctions,* V_T is the thermal voltage, V_{EB} is the emitter-base voltage, V_{CB} is the collector-base voltage and V_{AF} is the forward Early voltage [3]. This equation assumes that, when in reverse mode, the carriers injected from the collector reach either the emitter or the base. However, it was already noted that a part of the carriers injected by the collector are collected by the substrate. Thus, the $\exp\left(\frac{V_{CB}}{V_T}\right)$-term of this equation is not complete.

Assume the transistor is in forward mode and the emitter injects carriers which are collected by a *part* of the collector; let's call this part A_{eff}, i.e. the effective collector area. When for the emitter-base junction the saturation current is given by I_S, the saturation current of this effective area of the collector equals, according to the Ebers-Moll reciprocity condition [42] and [45], I_S also. Thus for the total saturation current of the base-collector junction yields:

$$I_{S_{BC}} = I_S \cdot \frac{A_{col}}{A_{eff}}, \qquad (6.89)$$

where A_{col} is the physical area of the collector-base junction. For a transistor without a substrate, the total collector area [4] is effective and the fraction in equation (6.89) reduces to one. Using this expression for the saturation current of the base-collector junction and rewriting it yields (it should be noted that base currents are ignored):

$$I_C = I_S \exp\left(\frac{V_{EB}}{V_T}\right) \cdot \left[1 - \frac{A_{col}}{A_{eff}} \exp\left(\frac{V_{CE}}{V_T}\right) \right] \cdot \left[1 + \frac{V_{BC}}{V_{AF}} \right]. \qquad (6.90)$$

Thus the least influence of the saturation effect is found when the total collector area is effective; when A_{col}/A_{eff} is one.

This expression is validated by measurements performed on two types of lateral transistors, a small one and a large one, see figure 6.18. In figure 6.19, equation (6.90) is depicted for the two transistors including the measured data. It is clear that the measured data fit the calculated expression very well. The value for A_{col}/A_{eff}, used for the calculated saturation behavior, was derived from $I_{S_{BC}}$ and $I_{S_{BE}}$ according to:

$$\frac{A_{col}}{A_{eff}} = \frac{I_{S_{BC}}}{I_{S_{BE}}}, \qquad (6.91)$$

[3] The reverse Early effect is ignored as it has no influence on the current finally obtained from the current sources, which are very often realized by current-copier-like structures.

[4] The total collector area is only the area of the collector that is contiguous with the base.

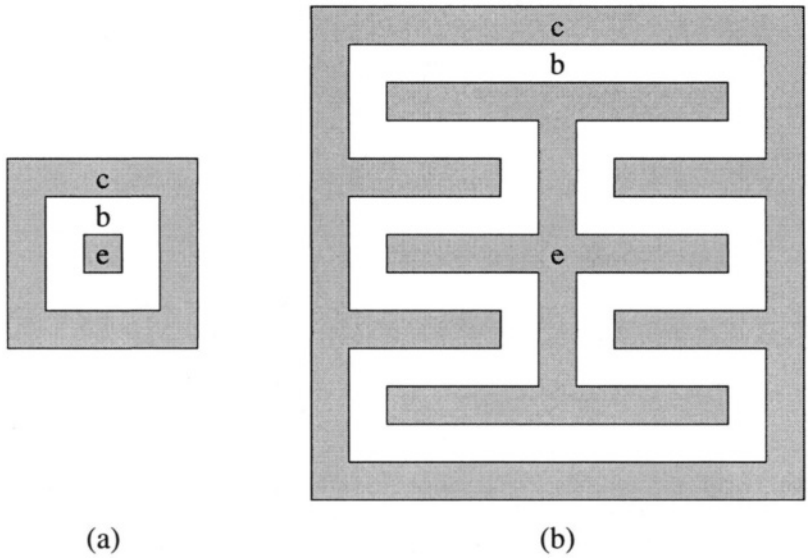

Figure 6.18: Two lateral PNP transistors. a) a small one with $A_{col}/A_{eff} \approx 2$ and b) a large one with $A_{col}/A_{eff} \approx 1$.

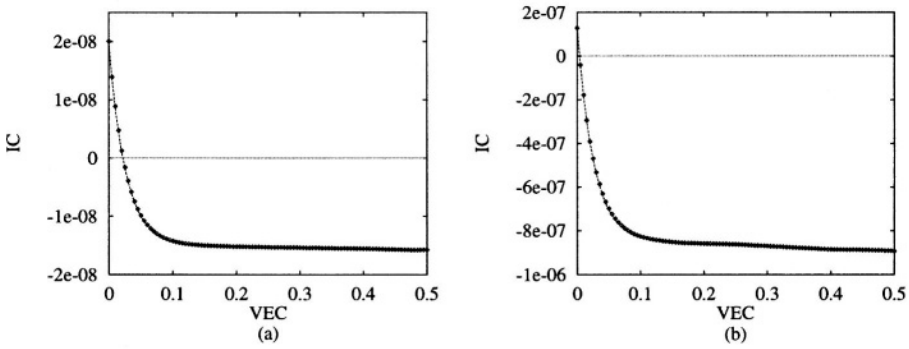

Figure 6.19: Plot of the calculated (solid line) and measured (dotted line) saturation behavior of a) the small and b) the large lateral PNP transistor.

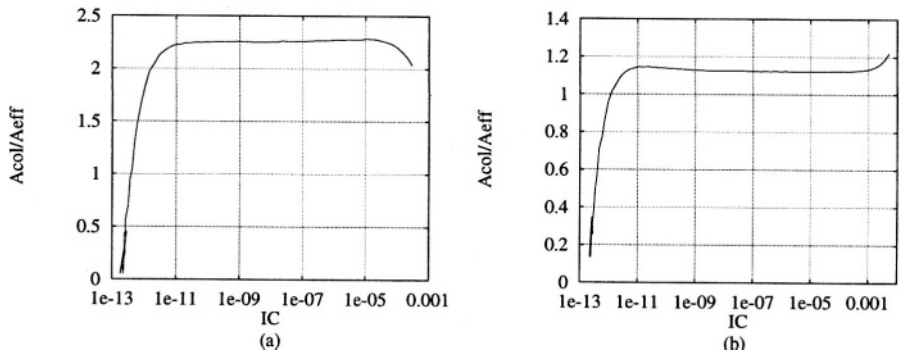

Figure 6.20: A_{col}/A_{eff} measured as a function of the collector current of the non-saturated transistor for a) the small PNP transistors and b) for the large one.

which were separately extracted from two Gummel plots; one plot for the normal forward transistor and one for the reverse transistor. The value for the forward Early voltage was extracted from the measured saturation behavior. In order to check the extent to which the ratio of the total collector area and effective collector area depends on the current level, the collector current at $V_{CE} = 0$ V was measured as a function of the collector current of the non-saturating transistor [5]. The measurements are depicted in figure 6.20. In the range from a few pA to several hundreds of μA, the ratio of the total collector area and the effective area is more or less constant. This is due to the fact that it is a relative quantity. For higher currents the bulk resistances and high-level effects influence this ratio, whereas for the lower currents the leakage currents in the measurement setup cause the drop of the plotted function.

Compared with the model as described in [46] the ratio A_{col}/A_{eff} corresponds to:

$$\frac{A_{col}}{A_{eff}} = X_{CS} + 1. \tag{6.92}$$

The parameter X_{CS} is not given any physical meaning there. From the discussions in this section X_{CS} can thus be seen as the ratio of the ineffective and effective collector-base junction area.

6.7.2.2 The vertical transistor

A typical structure for a vertical NPN transistor is given in figure 6.21. For this type of transistor the parasitic PNP transistor (base-collector-substrate) conducts when the NPN transistor saturates. The currents constituting the

[5] For $V_{CE} = 0$ V, this ratio can be very easily extracted from the collector current, see equation (6.90) in which the exponential function then becomes one.

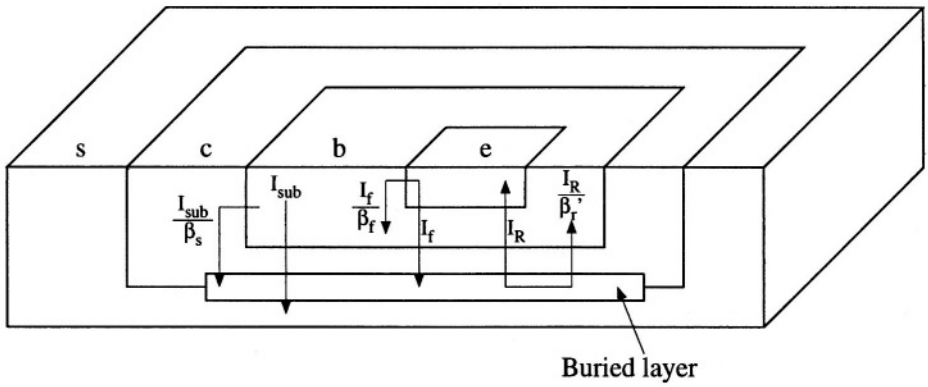

Figure 6.21: A typical structure for a vertical NPN transistor.

collector current are depicted in figure 6.21. The collector current, I_C, is now given by (see for instance [46]):

$$I_C = I_F - \left(1 + \frac{1}{\beta_r'}\right) I_R + \frac{I_{SUB}}{\beta_s}, \tag{6.93}$$

where I_F is the ideal forward-collector current, I_R is the ideal reverse-collector current, I_{SUB} is the substrate current (the "collector" current of the substrate transistor) and β_r' and β_s are the current-gain factors of the inverse transistor and the substrate transistor, respectively. For an increasing level of saturation, I_F is more compensated by I_R, and is completely canceled when the collector-emitter voltage is zero. This is a result of the equal saturation currents, I_S, of the base-emitter and base-collector junction (the Ebers-Moll reciprocity condition). For $V_{CE} = 0$ V, the base-emitter and base-collector voltages are equal and thus their injection currents are equal, I_F and I_R, respectively. Then only the two "base currents" remain of which the base current of the inverse transistor is dominant; β_s may be smaller than β_r', however due to the high doping of the buried layer, i.e. the base region of the substrate transistor, I_{SUB} can be considerably smaller than I_R and as a result the term due to the substrate current can be ignored. Then the collector current as a function of the collector-emitter voltage, including the forward Early effect, can be written as:

$$I_C = I_S \exp\left(\frac{V_{BE}}{V_T}\right)\left[1 - \left(1 + \frac{1}{\beta_r'}\right)\exp\left(\frac{V_{EC}}{V_T}\right)\right], \tag{6.94}$$

in which the expression for I_F and I_R are substituted. The value of β_r' can be on the order of 10 for minimally sized transistors. From this expression it follows that the designer can alter the saturation behavior only slightly by altering the base current of the inverse mode transistor (collector and emitter interchanged).

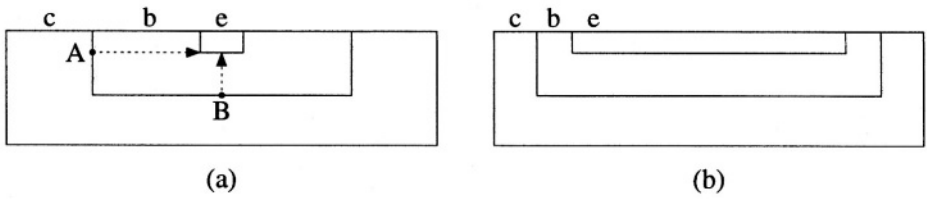

Figure 6.22: The influence of the layout on the reverse current-gain factor of a vertical transistor, a) a transistor with a relatively tiny emitter and b) a transistor with a relatively large emitter.

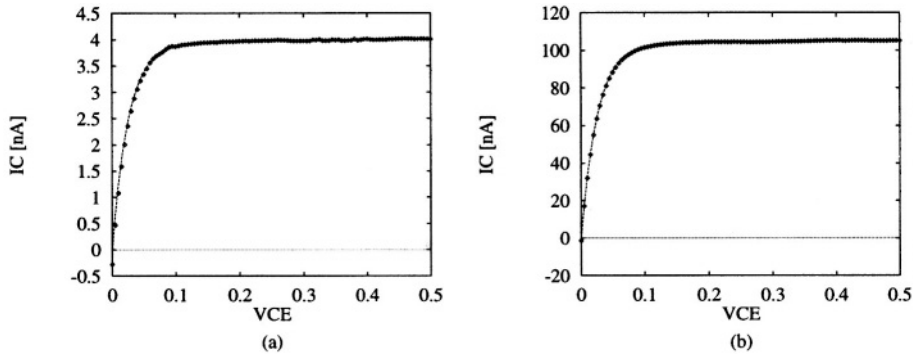

Figure 6.23: The calculated (drawn line) and the measured (dotted line) saturation behavior for a) a transistor with a relatively small emitter (6x6 μm^2) and b) a relatively large transistor (30x30 μm^2)

The influence of the layout on the reverse current-gain factor is depicted in figure 6.22, see also [35]. In figure 6.22a, a tiny emitter is made in a relatively large base and collector. When this transistor is in reverse mode, the collector functions as the reverse emitter. Many carriers injected from this reverse emitter have to travel a relatively large distance to the reverse collector. The carriers injected at point A have to travel a larger distance than the carriers injected at point B. The carriers injected at point A have a higher chance of recombining in the base compared to carriers injected at point B. Therefore, the transistor in figure 6.22b will have a lower base current (when the recombination is assumed to be the dominant effect) as the most of the injected carriers only have to travel a relatively short distance. Thus, to obtain a high reverse-current-gain factor, the emitter and collector must have as much overlap as possible (cf. the situation for the lateral transistor, in that case the overlap also had to be made as large as possible).

In figure 6.23, the measured and calculated saturation behavior for two vertical NPN transistors are depicted. The main difference between the two plots

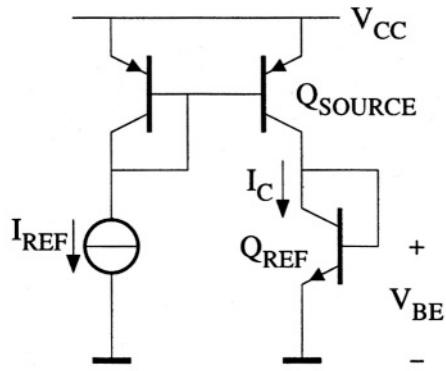

Figure 6.24: A simplified base-emitter voltage generator.

(besides the current level), is the relative value of the current at $V_{CE} = 0$ V. For the large transistor this current is relatively closer to zero than for the small transistor, which indicates a higher β_r'. In the calculated data, β_r' was assumed to be infinite. Calculating the reverse-current-gain factors from the measured data, they appeared to be 14 and 76 for the small and large transistor, respectively.

6.7.2.3 Influence on the reference voltage

The expressions found in the two previous sections for the collector current of a transistor when it is saturating can both be described by:

$$I_C = I_S \exp\left(\frac{V_{BE}}{V_T}\right)\left[1 - K\exp\left(\frac{V_{EC}}{V_T}\right)\right] \cdot \left[1 + \frac{V_{CB}}{V_{AF}}\right], \qquad (6.95)$$

in which the polarities are chosen for an NPN type of transistor. For the vertical transistor K equals $1 + \frac{1}{\beta_r'}$ whereas for the lateral transistor K equals A_{col}/A_{eff}. To find the influence on the reference voltage of this saturation, the simplified base-emitter voltage generator of figure 6.24 is studied. For relatively low temperatures, base-emitter voltages tend to become relatively large. As a consequence, the current source may saturate and an error is found in the base-emitter voltage V_{BE}. This error depends on the level of saturation. As the saturation of the current source is a second-order effect, in the sense of not being a key parameter, only the first-order temperature behavior of the base-emitter voltage of Q_{REF} has to be taken into account. In that case the collector-base voltage of Q_{SOURCE} varies linearly with the temperature and its influence on the current via the Early effect can consequently neatly be compensated, if necessary. Further, by choosing a transistor with a relatively high forward Early

voltage, this effect can be made negligible.

The forward Early voltage itself is also slightly temperature dependent as the base-collector voltage is not large compared with the build-in potential [47]. However, the resulting error is also negligibly small compared with the error due to the saturation. Therefore, for calculating the influence on the collector current when the transistor is saturating, the last factor of equation (6.95) can be ignored. In that case the base-emitter voltage of Q_{REF} can be written as:

$$V'_{BE} = V_T \ln \left\{ I_S \exp\left(\frac{V_{BE}}{V_T}\right) \cdot \left[1 - K \exp\left(\frac{V_{EC}}{V_T}\right)\right] \right\}, \tag{6.96}$$

where V'_{BE} is the base-emitter voltage including the error due to saturation of the current source. This expression can be rewritten as:

$$V'_{BE} = V_{BE} + V_T \ln \left[1 - K \exp\left(\frac{V_{EC}}{V_T}\right)\right]. \tag{6.97}$$

The second term on the right-hand side of this expression is due to the saturation of the current source. This error should be relatively small and can therefore be approximated by:

$$V_{ERROR} = -V_T \cdot K \exp\left(\frac{V_{EC}}{V_T}\right). \tag{6.98}$$

From this expression the minimum voltage required for a current source can be found.

> Example: Assume the error in a base-emitter voltage (say 700 mV) due to a saturating current is allowed to be 100ppm.
>
> Question: What is the minimum voltage required for the current source (take 273 K for the ambient temperature)?
>
> Answer: V_{ERROR} equals 100ppm times 700 mV which is 70 μV. From expression (6.98) the minimum voltage required is found to be: $V_{CE} = 140$ mV. Thus for a supply voltage of 1 V at least about 15 % is required for the current source.

6.8 Special structures

Special cases of bandgap references can be found. These arise when the circuits are reduced to having only one scaling factor. In other words, these cases are found when one of the scaling factors is chosen to be one.

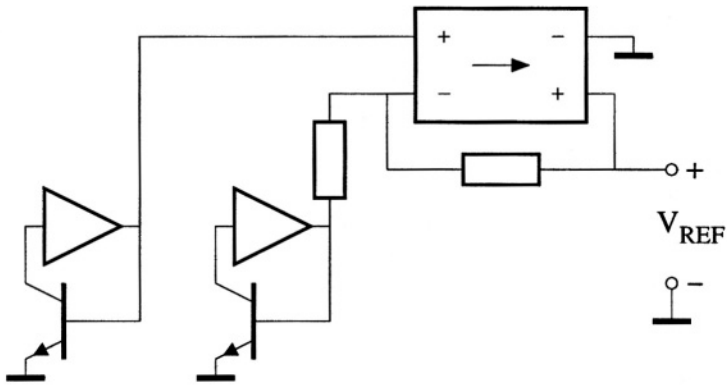

Figure 6.25: The first-order compensated bandgap reference with output voltage equal to $V'_G(0)$.

6.8.1 First-order compensated reference

When for a first-order compensated bandgap reference the reference voltage is chosen to be $V'_G(0)$, the sum of the two scaling factors becomes one, see equation (6.27), or, in other words, the two scaling factors differ by one. As a result a simplification of the circuit can be made, see figure 6.25. The bandgap reference as proposed by Brokaw [8] is found. To obtain a correctly compensated reference voltage, equations (6.26) and (6.27) have to be fulfilled (the constant term equal to V_{REF} and the first-order term zero). As a result of the topology, the sum of the two scaling factors is always one, and thus the two first-order terms of the base-emitter voltages always compensate for each other when the output voltage is tuned to $V'_G(0)$. The bandgap reference can therefore be trimmed by means of trimming a bias current or the scaling factor at one temperature. When non-idealities introduce some additional first-order temperature dependencies in the bandgap reference, the value to which the reference voltage must be tuned in order to obtain first-order compensation changes.

For the references with two separate scaling factors, it is preferable to tune the base-emitter voltages for trimming the bandgap reference, over tuning one or more scaling factors. As the scaling factors are realized by means of resistor ratios, these are likely to be more accurately determined than the absolute values of the emitter areas. When the reference is accurately designed, only process variations need to be compensated for and those are becoming better due to the fact that a transistor with a given emitter size becomes a relatively large transistor when realized in a sub-micron technology. When unexpected (relatively large) errors are found, one can always use one of the scaling factors to compensate for this.

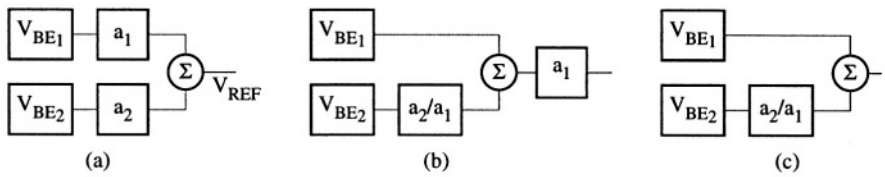

Figure 6.26: Simplification of a second-order compensated bandgap reference.

6.8.2 Second-order compensated reference

When one of the scaling factors is shifted to the output of a second-order compensated bandgap reference, a simplification can be made. Consider the general bandgap reference as depicted in figure 6.26a. Of this bandgap reference, scaling factor a_1 is shifted through the summing node, see figure 6.26b. The output voltage of the reference is still the original value. As the scaling factors are assumed to be temperature independent, the input voltage of scaler a_1 in figure 6.26b, is also temperature compensated. Therefore, assuming a_1 to be one, only one scaling factor remains and still a temperature compensated reference voltage is obtained, see figure 6.26c. As, however, one degree of freedom is used by assuming a_1 to be one, the reference voltage can no longer be freely chosen.

In figure 6.26, a_1 is shifted through the summing node as this is the positive scaling factor. Shifting a_2 through the summing node and assuming it to be one, results in a negative reference voltage; to obtain a positive reference voltage $|a_2|$ has to be shifted through the summing node.

The reference voltage that is found when a_1 is assumed to be one follows directly from equation (6.40), yielding:

$$V_{REF} = V'_{G1}(0) - V'_{G2}(0) \cdot \frac{B_2(\theta_1)}{B_2(\theta_2)}. \tag{6.99}$$

This reference voltage only depends on the two θs and process parameters. For the θs being 1 and 0, and the model of Varshni for the bandgap energy [30], a reference voltage of:

$$V_{REF} = 245mV, \tag{6.100}$$

is found. The new scaling factor $a'_2 = a_2/a_1$ is completely responsible for the second-order compensation (the output voltage of the reference of which $a_2 = 1$ equals -312 mV). As a_2 is realized by a ratio of components, a_2 can be accurately realized on a chip. Thus second-order compensation is readily achieved. Subsequently, to obtain the first-order compensation, the output voltage only has to be trimmed to 245 mV. Thus trimming at only one temperature, in order to obtain a second-order compensated reference voltage, is sufficient.

6.9 Conclusions

In this chapter a structured design method for bandgap references has been presented. The chapter started with a historical overview of the bandgap references being designed and published over the last thirty years. From this overview the conclusion was drawn that most of the bandgap references have in common the fact that the temperature behavior of one base-emitter voltage is compensated by the temperature behavior of one or more other base-emitter voltages. Some publications deal with MOS-realizations and others describe the use of the temperature behavior of the base current for a temperature compensation of a base-emitter voltage. But the most accurate compensation is obtained when the underlying physical effect is the same for the compensated and compensation voltage.

In order to get a general description of the principal bandgap reference function, it was described in this chapter as a linear combination of base-emitter voltages. This linear combination was described using the Taylor series of the base-emitter voltages. Subsequently, the constraints of an nth-order temperature compensation was introduced, resulting in a set of equations which had to be solved. It appeared that for realizing an nth-order temperature compensation at least n, or when n is smaller than 2, two base-emitter voltages are required. Further, for compensating m higher orders (order ≥ 2), at least m collector currents with a different temperature dependency are required. Thus, for realizing a second-order compensated bandgap reference, only two base-emitter voltages with a different temperature dependency of the collector currents are required.

After the identification of the key parameters: the bandgap energy, E_G, the saturation current of the base-emitter junction, I_S, including its order of temperature dependency, η, and the reverse Early voltage, V_{AR}, a complete description is obtained of the core components of the bandgap reference and the design with respects to the three fundamental aspects can be done. This design should preferably be done orthogonally.

For the design of the noise behavior of the bandgap references, a distinction was made between the idealized bandgap reference (i.e. scalers, adders and bias ideal) and the practical implementation.

For the first-order compensated idealized bandgap reference, realized with the minimum of two base-emitter voltages, it was found that for a minimum noise level an optimum ratio of the two collector currents exists which only depends on the ratio of the two saturation currents. For an increasing ratio of these two saturation currents, the noise level reduces and the function, noise level versus ratio of the two collector currents, flattens. As a result of this flattening the noise behavior of the scaling factors could be minimized more or less orthogonal to the noise minimization of the idealized bandgap reference.

For the second-order compensated bandgap reference, also realized with the minimum number of two base-emitter voltages required, the minimum noise level

is found for a collector-current ratio which only depends on the second-order temperature dependencies of the base-emitter voltages. The noise level appeared to be dependent on a constant with only process parameters and the current consumption. Further, for a given current consumption the signal-to-noise ratio of a second-order compensated idealized bandgap reference is fixed, i.e. it is independent of the reference voltage. As the scaling factors are completely determined by the temperature compensation for the second-order compensated bandgap reference, the noise minimization of these scalers concerns only the division of the total current available for them between the two scalers. This minimum also appeared to be independent of the current consumption.

Thus both the noise level of the idealized bandgap reference and the scalers can be minimized independently. As their final noise level is inversely proportional to the current consumption, a global minimum is found when the current is optimally divided between the scalers and the idealized bandgap reference.

The noise production of the bias sources can be accounted for by doubling the noise sources from the active devices which were found up to now. Further, for the noise injected from the power supply via the finite output impedances of the current sources, resulting in extra noise at the output of the bandgap reference, an expression was derived giving the maximum attainable PSRR, assuming that the Early voltages of the current sources for the transistors in the idealized bandgap reference are equal.

The bandwidth of a bandgap reference is not limited to only DC; for higher frequencies the bandgap reference still has to behave well. This means that its output impedance has to remain relatively low which can be realized by means of an active output. Further, the output impedance of the bias current sources reduce for higher frequencies and consequently more noise is injected from the power supply. The PSRR can be increased again for higher frequencies by means of cancellation of the injected noise, which relies on realizing an appropriate ratio of the output impedances of the currents sources. But increasing the PSRR is also possible via suppression/compensation by making the scaling factors of the linear combination frequency dependent. Of course, a low-pass filter after the bandgap reference could also do the job, however, very likely at the cost of a larger capacitance.

The signal power of the bandgap reference is given in the specification of the output voltage. The efficiency of the supply of this signal power can be increased by lowering the supply voltage to the reference voltage. This inherently means a reduction of the voltage available for the current sources. These current sources must be prevented from saturation and for this purpose an expression was derived relating the error in a base-emitter voltage of the idealized bandgap reference to the maximum voltage for the current sources available.

Finally, two special structures were discussed. For one of the structures the second-order temperature compensation depends on the matching of resistors

only. Thus, a good second-order temperature compensation without trimming is easily realized.

Bibliography

[1] R.J. van de Plassche. *Integrated Analog-to-Analog and Digital-to-Analog Converters*. Kluwer, Boston, 1994.

[2] M.M. Martins and J.A.S. Dias. CMOS shunt regulator with bandgap reference for automotive environment. In *IEE Proceeding Circuits Devices and Systems*, volume 141, pages 157–161, June 1994.

[3] H. Tanaka et al. Sub-1-μA dynamic reference voltage generator for battery-operated DRAMs. *IEEE Journal of Solid-State Circuits*, 29(4):448–453, April 1994.

[4] D.F. Hilbiber. A new semiconductor voltage standard. In *ISSCC Digest Technical Papers*, volume 7, pages 32–33, 1964.

[5] R.J. Widlar. New developments in IC voltage regulators. *IEEE Journal of Solid-State Circuits*, 6(l):2–7, February 1971.

[6] R.J. Widlar. Some circuit design techniques for linear integrated circuits. *IEEE Transactions on Circuit Theory*, 12(4):586–590, December 1965.

[7] K.E. Kuijk. A precision reference voltage source. *IEEE Journal of Solid-State Circuits*, 8(3):222–226, June 1973.

[8] A.P. Brokaw. A simple three-terminal IC bandgap reference. *IEEE Journal of Solid-State Circuits*, 9(6):388–393, December 1974.

[9] G.C.M. Meijer and J.B. Verhoeff. An integrated bandgap reference. *IEEE Journal of Solid-State Circuits*, 11:403–406, June 1976.

[10] R.J. Widlar. Low voltage techniques. *IEEE Journal of Solid-State Circuits*, 13(6):838–846, December 1978.

[11] R.J. Widlar. A new breed of linear ICs runs at 1-volt levels. *Electronics*, pages 115–119, March 29 1979.

[12] C.R. Dobkin and R.C. Palmer. A curvature corrected micropower voltage reference. In *ISSCC Digest Technical Papers*, pages 58–59, February 1981.

[13] G.C.M. Meijer, P.C. Schmale, and K. van Zalinge. A new curvature-corrected bandgap reference. *IEEE Journal of Solid State Circuits*, 17(6):1139–1143, December 1982.

[14] I. Lee, G. Kim, and W. Kim. Exponential curvature-compensated BiCMOS bandgap references. *IEEE Journal of Solid-State Circuits*, 29(11):1396–1403, November 1994.

[15] E.A. Vittoz and J. Fellrath. CMOS analog integrated circuits based on weak inversion operation. *IEEE Journal of Solid-State Circuits*, 12(3):224–231, June 1977.

[16] E.A. Vittoz and O. Neyroud. A low-voltage CMOS bandgap reference. *IEEE Journal of Solid-State Circuits*, 14(3):573–577, June 1979.

[17] G. Tzanateas, C.A.T. Salama, and Y.P. Tsividis. A CMOS bandgap voltage reference. *IEEE Journal of Solid-State Circuits*, 14(3):655–657, June 1979.

[18] B.S. Song and P.R. Gray. A precision curvature-compensated CMOS bandgap reference. *IEEE Journal of Solid-State Circuits*, 18(6):634–643, December 1983.

[19] S.L. Lin and C.A.T. Salama. A $V_{be}(T)$ model with application to bandgap reference design. *IEEE Journal of Solid-State Circuits*, 20(6):1283–1285, December 1985.

[20] O. Salminen and K. Halonen. The higher order temperature compensation of bandgap voltage references. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 10–13, May 1992.

[21] E.A. Vittoz. MOS transistors operated in the lateral bipolar mode and their application in CMOS technology. *IEEE Journal of Solid-State Circuits*, 18(3):273–279, June 1983.

[22] M.G.R. Degrauwe et al. CMOS voltage references using lateral bipolar transistors. *IEEE Journal of Solid-State Circuits*, 20(6):1151–1156, December 1985.

[23] H.J. Oguey and B. Gerber. MOS voltage reference based on polysilicon gate work function difference. *IEEE Journal of Solid-State Circuits*, 15(3):264–269, June 1980.

[24] G.C.M. Meijer. Bandgap references. In J.H. Huijsing et al., editors, *Analog Circuit Design*, pages 243–268. Kluwer, Dordrecht, 1995.

[25] S.M. Sze. *Physics of Semiconductor Devices.* John Wiley & Sons, New York, 1969.

[26] Y.P Tsividis. Accurate analysis of temperature effects in $I_C - V_{BE}$ characteristics with application to bandgap reference sources. *IEEE Journal of Solid-State Circuits*, 15(6):1076–1084, December 1980.

[27] K. v.d. Lingen. *Bipolar Transistors for usage in Monolithic Bandgap References and Temperature Transducers.* PhD thesis, Delft University of Technology, Delft, 1996.

[28] J.W. Slotboom and H.C. de Graaf. Bandgap narrowing in silicon bipolar transistors. *Solid-State Electronics*, 19:857–862, October 1976.

[29] B. Gilbert. Unusual voltage and current references for IC realization. In J.H. Huijsing et al., editors, *Analog Circuit Design*, pages 268–352. Kluwer, Dordrecht, 1995.

[30] Y.P. Varshni. Temperature dependence of the energy gap in semiconductors. *Physica*, 34:149–154, 1967.

[31] A. van Staveren, J. van Velzen, C.J.M. Verhoeven, and A.H.M. van Roermund. An integratable second-order compensated bandgap reference for 1 V supply. *Analog Integrated Circuits and Signal Processing*, 8:69–81, 1995.

[32] I.E. Getrue. *Modeling the Bipolar Transistor.* Elsevier, New York, 1978.

[33] MicroSim Corporation. *Manual PSPICE 4.05.*

[34] L.K. Nanver, E.J.G. Goudena, and H.W. van Zeijl. DIMES-01, a baseline BIFET process for smart sensor experimentation. *Sensors and Actuators Physical*, 36(2):139–147, 1993.

[35] A. van Staveren. Chapter 5, Integrable DC sources and references. In W.A. Serdijn, C.J.M. Verhoeven, and A.H.M. van Roermund, editors, *Analog IC Techniques for Low-Voltage Low-Power Electronics.* Delft University Press, 1995.

[36] A. van Staveren, C.J.M. Verhoeven, and A.H.M. van Roermund. The influence of the reverse early effect on the performance of bandgap references. *IEEE Transactions on Circuits and Systems I,* 43(5):418–421, May 1996.

[37] V.I. Anisimov et al. Circuit design of low-power reference voltage sources. *Telecommunications and radio engineering, Part 1*, 48(1):11–17, 1993.

[38] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[39] A. van Staveren, C.J.M. Verhoeven, and A.H.M. van Roermund. The design of low-noise bandgap references. *IEEE Transactions on Circuits and Systems I*, 43(4):290–300, April 1996.

[40] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

[41] A. Bilotti and E. Mariani. Noise characteristics of current mirror sinks/sources. *IEEE Journal of Solid-State Circuits*, 10(6):516–524, December 1975.

[42] J.J. Ebers and J.L. Moll. Large-signal behavior of junction transistors. In *Proceedings of the I.R.E.*, volume 42, pages 1761–1772, December 1954.

[43] H.C. Lin et al. Lateral complementary transistor structure for the simultaneous fabrication of functional blocks. *IEEE Proceedings*, 52:1491–1495, December 1964.

[44] J. Lindmayer and W. Schneider. Theory of lateral transistors. *Solid-State Electronics*, 10:225–234, 1967.

[45] B.L. Hart. Direct verification of the Ebers-Moll reciprocity condition. *International Journal of Electronics*, 31(3):293–295, 1971.

[46] H.C. de Graaff and F.M. Klaassen. *Compact Transistor Modeling for Circuit Design*. Springer-Verlag, Wien, 1990.

[47] W.C. Dillard and R.C. Jaeger. The temperature dependence of the amplification factor of bipolar-junction transistors. *IEEE Transactions on Electron Devices*, 34(1):139–142, January 1987.

Chapter 7

Conclusions

The mathematical description language, and especially the differential equation, is a good and objective language for describing the signal-processing function to be implemented. When orthogonalization with respect to noise, signal power and bandwidth is a key item in the structured design method, relatively fast and cheap design trajectories can be obtained.

The main drawback of low-voltage, low-power design is the performance degradation of a design resulting from the inherent low-current constraint. Reducing the maximum allowed current consumption of a circuit results in a relatively large reduction of the information-handling capability; the quality with respect to noise, signal power and bandwidth reduces.

Functions which are still ideal at the level of differential equations, and thus intrinsically have a high performance, are the harmonic frequency reference (harmonic oscillator) and the constant reference (bandgap reference). These functions are found from the homogeneous linear differential equations. An additional important function for designing the corresponding implementations is the scaler (amplifier).

The sum of the loop poles, related to the LP product, is a measure for the maximum number of dominant poles, which is an important figure of merit for potential frequency behavior.

Further, clipping distortion of a negative-feedback amplifier is best prevented by taking care that there is enough signal-driving capability whereas the g_m and the β distortion are best reduced by increasing the overall loop gain.

Noise is a key issue in the structured design of high-performance harmonic oscillators. Tapping a resonator results in a relatively large improvement of the CNR of the complete oscillator as a result of a much better noise match. This tapping does not necessarily degrade the intrinsic performance of the resonator.

The tap factor should be relatively low compared with the quality factor of the resonator to be tapped. This limitation is a result of the weaker coupling

between the active part and the resonator for larger tap factors.

A linear combination of junction voltages is a good approach for synthesis of bandgap references. From that description it follows that with two scaled base-emitter voltages with collector currents having a different temperature dependency, a temperature compensation can be obtained up to the second order.

The remaining temperature dependency of a bandgap reference as a result of limited compensation is fundamentally limited by the temperature dependency of the injection of carriers for a given junction voltage, which includes the temperature dependency of the bandgap energy. Practical limits are found in the temperature dependency of resistors and the saturation of current sources.

The absolute noise level of the first and second-order compensated bandgap references realized by the minimum number of base-emitter voltages (two) is approximately inversely proportional to the sum of the two corresponding collector currents. The constraint for an optimum relative-noise-level, is independent of the sum of collector currents but, in contrast, poses conditions for the ratio of the two collector currents.

The ratio of these two collector currents is given for the optimum relative noise level; it is independent of their sum.

Chapter 8

Summary

The increasing complexity of electronic systems calls for structured design methods in order to obtain optimal performance for minimal design time and design cost, (like chip area and power dissipation, et cetera).

The structured design method as presented in Chapter 2 shows the mathematical description language, i.e. differential equations, to be an appropriate language for specifying the required signal-processing function. A key issue of the method presented is orthogonalization; the three fundamental aspects of the information-handling capability:

- noise;
- signal power;
- bandwidth;

are optimized in subsequent, ideally independent, design steps. As a result, for each of these three aspects the corresponding fundamental limits can be derived at a relatively early stage in the design process. When the corresponding design requirements are beyond the fundamental limits, no solution can be found. The practical limits, i.e. the limits introduced by technology, for instance, should be between the design requirements and the fundamental limits.

From the simplest differential equations, the homogeneous linear differential equations, the following functions are found:

- harmonic frequency reference;
- constant reference.

These are derived from the second and first-order homogeneous linear differential equations, respectively. No fundamental obstacles prevent these functions from

being ideal. Therefore, these functions should be used when a high-performance frequency reference or a DC reference is required, respectively. Further, from an implementation point of view, the scaler is an inevitable function when implementing those references.

Low-voltage, low-power design becomes more and more necessary for realizing portable, high-functional equipment. Chapter 3 dealt with low-voltage, low-power design with respect to the two orthogonal design aspects:

- low voltage;
- low current.

A low-voltage constraint mainly has an effect on the maximum signal voltages, or signal currents which are related via impedance levels. As active devices have an expanding function from input voltage to output current, and since the voltage is generally given for power supplies and the current is free, current-mode design is a perfect solution for coping with a low-voltage constraint.

A low-current constraint is shown to have a tremendous effect on the information-handling capability of electronics. Independent of the signal type, i.e. voltage or current, the performance with respect to noise, signal power and bandwidth reduces for lower currents.

Low-power design is a combination of low-voltage and low-current design, and, therefore, the corresponding effects of both apply. On top of that, stating a low-power constraint hampers the orthogonality of a design process.

In the subsequent chapters, structured electronic design of the sealer (Chapter 4), the harmonic frequency reference (Chapter 5) and the constant reference (Chapter 6) are discussed.

An implementation of a scaler is the amplifier. Chapter 4 shows the feasibility of the orthogonalization of noise, signal power and bandwidth in the design of amplifiers.

Noise is mainly found at the input of the amplifier and depends on the source impedance and the input device. Once minimized, the noise of the amplifier can no longer improve.

The maximum signal power of an amplifier is limited by distortion. Distinction is made between clipping distortion and weak distortion. Clipping distortion has to be prevented by ensuring that signal voltages/currents remain within the available supply voltage/current. Weak distortion, i.e. β and g_m distortion, can effectively be reduced by an increased loop gain of the overall feedback loop; local feedback does not result in an improvement.

Bandwidth is shown to be the concern of the complete feedback loop. The product of the dominant loop poles and the DC loop gain, the LP product, is a measure of the absolute frequency behavior. The dominant loop poles are the largest set of loop poles for which it holds that the sum of loop poles is less negative than the sum of the corresponding system poles.

With frequency-compensation techniques the relative frequency behavior can be designed. Depending on the type of technique, the LP product is more or less reduced. Further, frequency-compensation techniques using local feedback are preferred to techniques without feedback. The latter degrades the distortion characteristics of an amplifier whereas the former has hardly any effect.

The harmonic oscillator, discussed in Chapter 5, is an implementation of the harmonic frequency reference. The differential equation of its core, a resonator, closely resembles the differential equation of the ideal harmonic frequency reference. By means of an active undamping, they can be made equal. To obtain the intrinsic high performance of the resonator, the influence of the active undamping should be negligibly small.

For maximum signal power, the series resonator is shown to be preferred to the parallel resonator; for a given power supply voltage, the power of the series resonator can be a factor Q (the quality factor of the resonator) higher than the power of the parallel resonator.

Noise is shown to be a major problem in harmonic oscillators. Due to the extreme values of the impedance of the resonator at resonance, relatively low for the series resonator and relatively high for the parallel resonator, noise of the active part has a tremendous effect on the overall noise performance. By means of tapping the impedance level of the resonator can be changed, up to a certain extent, without degrading its intrinsic high performance. Due to this impedance transformation a much better noise match can be obtained. With a design example of a low-voltage harmonic oscillator an improvement was shown of about 14 dB for equal power consumption.

The maximum tap factor is limited by the degradation of the intrinsic quality factor. This occurs for tap factors on the order of, or higher than, the quality factor. This is mainly a result of the weaker coupling between the resonator and the undamping.

As a result of the tapping, an equivalent parallel capacitance is found at the resonator terminals. To prevent the oscillator from parasitic relatively high-frequency oscillations, the resonator should be undamped by a negative resistance *and* a negative capacitance.

The effect of a limited bandwidth of the undamping is shown to be a lower effective quality factor.

Finally, Chapter 6 describes the structured electronic design of bandgap references, i.e. an implementation of the constant reference. A bandgap reference can be described as a linear combination of base-emitter voltages. From the temperature behavior of the base-emitter voltage follows that to compensate n orders of the temperature dependency of a linear combination of base-emitter voltages, at least n, or when n is smaller than two, two base-emitter voltages are required, and for compensating m orders (order ≥ 2), at least m collector currents with a different temperature dependency are required.

The absolute minimum-noise-level at the output of the first and second-order compensated bandgap reference with a minimum number of base-emitter voltages (two), depends, approximately, inversely proportionally on the *sum* of the collector currents. From the constraint of a relative minimum-noise-level, conditions are found for the *ratio* of the two collector currents only. Further, by choosing the appropriate scaling of the emitter areas, the noise minimization of the scalers and adders can be made orthogonal to this minimization.

In contrast with the required signal which is only at DC, the bandgap reference still needs to have a relatively low output impedance for higher frequencies. The bandwidth of the output impedance can be designed orthogonally to the other design aspects.

The maximum signal of the bandgap reference has, on the one hand, to do with the current driving capability and on the other hand with the value of the reference voltage relative to the supply voltage.

The current driving capability is determined by the output stage of the bandgap reference and its biasing.

A relatively maximum signal voltage for a given reference voltage is obtained when the voltage supply is lowered such that it is relatively close to the reference voltage; this is in fact improving the power efficiency. This inherently means that the bias-current sources are close to saturation. Transistors showing the lowest degree of saturation are those which have an emitter and collector which overlap as much as possible. A relation is derived for the additional temperature dependency of a bandgap reference as a result of saturating current sources. From that it follows that for high-performance bandgap references, i.e. with dependencies of only a few hundred ppm over a temperature range of about 100 K, the minimum voltage required for current sources is about 150 mV.

In the appendix several design examples can be found. One design example deals with the design of a low-voltage, low-power first-order compensated bandgap reference for which the noise minimization is the main topic; calculations are in good agreement with simulations. An other example discussed the design of a low-voltage, low-power second-order compensated bandgap reference. For this design the compensation of the temperature dependency by means of a linear combination was the main topic. Theoretically the minimum temperature dependency over a temperature range of 100 K is about 22 ppm. The measurement results show a temperature dependency of about 150 ppm over this temperature range, but it is still the best 1 V bandgap reference found in literature. The additional temperature dependencies result from saturating current sources. By choosing processing in the current domain or by increasing the supply voltage even better results can be obtained.

Appendix A

Minimum voltage required for feedback amplifiers

To use nullors in a sensible way they must have some kind of feedback. As voltage and current can be the signal domain at the input and output of the nullor, four types of single-loop configurations can be distinguished, see figure A.1. From the figure it may become clear that when a port has a grounded terminal the single CE stage can be used having the lowest minimum required voltage, $V_{DEVICE}+V_{MARGIN}$. When a port is floating, the two inputs/outputs of a stage are needed. For a floating input port both the single and balanced CE stage can be used as the input stage. This is due to the fact that the input signal of a nullor implementation tends to zero. Of course, for DC this only holds for the balanced CE stage. For a floating output stage a balanced output stage is required. Were a single CE stage to be used, then due to the anti-phase relation between the two nullor outputs, the maximum available voltage swing would be considerably reduced. When one output increases the other decreases; for a single CE stage this means that when the emitter voltage increases, the collector voltage decreases which may result in saturation (NPN). If a balanced output is used, no signal is present at the common emitter node and the voltage swing is again only determined by the collector voltage swing.

When floating ports are required, they will have a somewhat larger minimum required voltage due to the fact that the port is connected between two current sources instead of one current source and ground. Using a stage with balancing in the current domain (when balancing is required [1]) is only possible when the nullor has one terminal grounded such that the common emitter node can be grounded and the port at which it is used remains floating. The first

[1]Balancing in the current domain may also be used instead of a single CE stage in order to make profitable use of the features of balancing (canceling of even-order distortion et cetera).

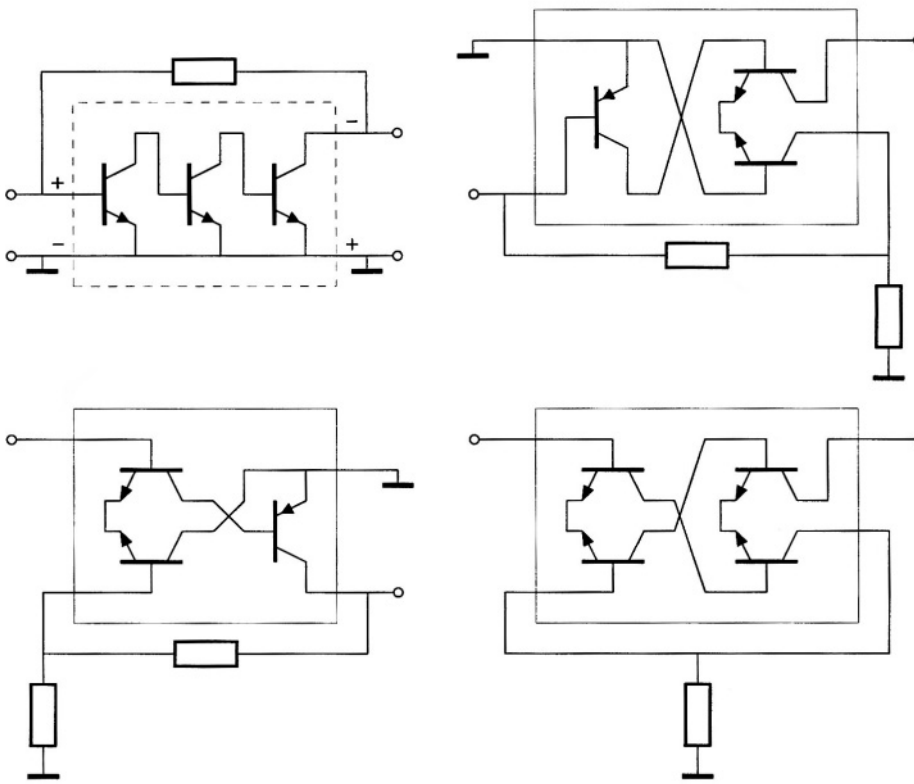

Figure A.1: The four possible single-loop amplifiers.

constraint states that it is only possible for the voltage and current amplifier; whereas the second constraint states that it can only be used at the output of a nullor implementation. Thus this stage can be profitably used in a current amplifier. This is elucidated by the two examples in figure A.2. For the current amplifier, the output stage can be replaced by a stage with balancing in the current domain as one input terminal of a current amplifier is grounded and the output of the amplifier remains floating. [2] For the voltage-to-current amplifier, only the two signal transistors at the input are depicted for the sake of clarity. The following problems occur. As the signal source is now in parallel with the input of an input device, the maximum voltage swing is now considerably reduced (for MOS(FETs) a considerable voltage swing may still be found). This is mainly caused by the fact that the port is no longer floating and a fifth terminal is introduced in the nullor implementation. Correct voltage comparison (maintaining the large input voltage swing) is only possible when an element is used in series with the nullator across which, by means of the action of a norator, the input voltage is canceled. Now this canceling takes place across the devices realizing the nullator; using indirect voltage comparison is no solution.

Summarizing, the transimpedance amplifier and the current amplifier can have the lowest minimum required supply voltage. The minimum required supply voltage for the voltage amplifier and the transconductance amplifier is somewhat larger, i.e. the saturation voltage of a current source.

[2] As the differential stage is asymmetrically driven, the other input signal for this stage is supplied by the control loop realizing the balance. Consequently, the bandwidth of the control loop must be at least as large as the bandwidth of the overall amplifier. In the case of a symmetric drive, the control loop only has to act on the mean value and thus may have a relatively low bandwidth.

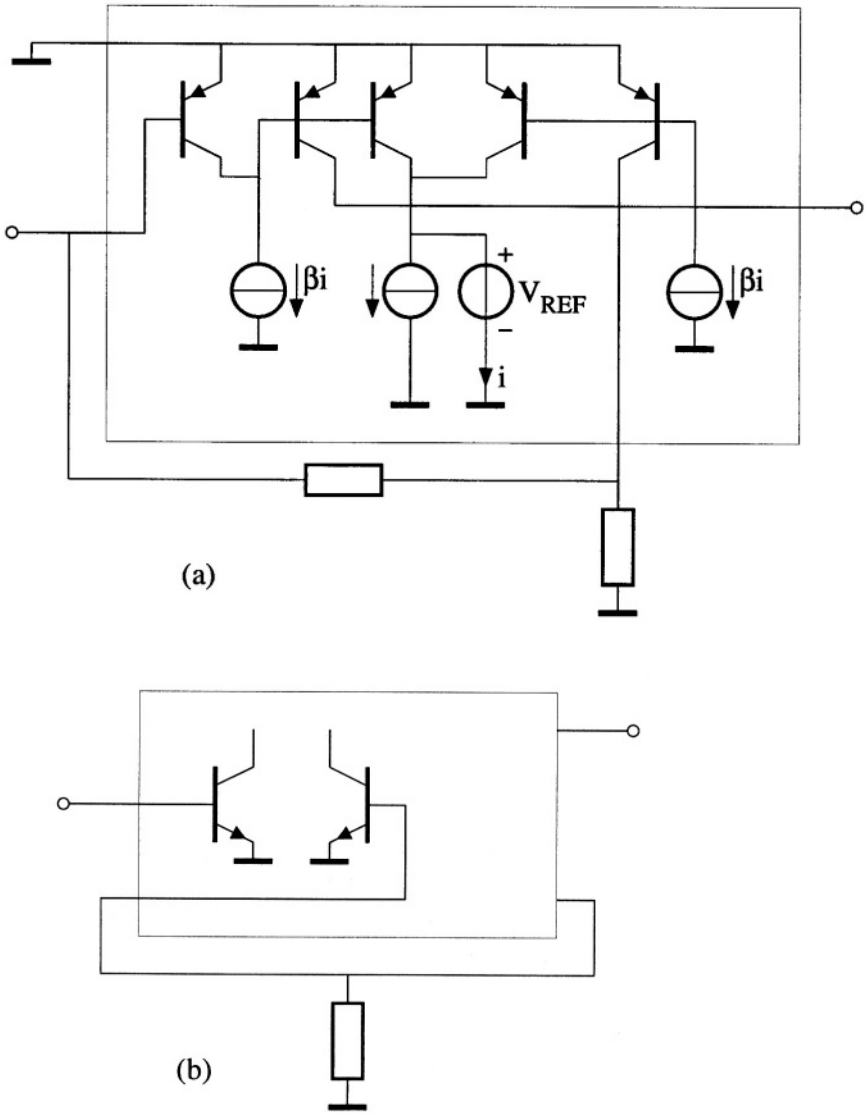

Figure A.2: Using a balanced stage in the current domain for a) a current amplifier and b) a voltage-to-current amplifier.

Appendix B

Design example: class-AB amplifier

B.1 Introduction

A current trend in electronic design, is the design of low-voltage (1 V) low-power (μW) circuits. These circuits have the advantage that they can be powered by a single, relatively small battery; single as the supply voltage is only 1 V and relatively small as the power consumption is low. The output amplifier presented [1] is a part of a completely-integrated single-chip 1 V LW receiver.

The effect of the low-voltage constraint (1 V) is mainly found in the circuits topology. For high supply voltages, components and function blocks can be stacked between the supply rails; for a 1 V supply they need to be placed in parallel to meet the low-voltage specification [2].

The low-power constraint demands that the efficiency of a circuit must be as close to 100% as possible, and all the power from the power supply must be directed to the load. For a given supply voltage (without the use of charge pumping techniques, etc.) the low-power constraint becomes a low-current constraint. All the current from the power supply must go through the load and all the other currents must ideally be zero. This constraint has the largest influence on those parts of the circuit where the signal currents are the largest. For instance, a small improvement in the efficiency of the output amplifier can result in a reduction in current consumption equal to the total current consumption of the input amplifier of the LW receiver.

Here the design of a highly efficient output amplifier for a 1 V supply is described. Section B.2 describes the specifications and the choice of the basic configuration of the amplifier. In this section nullors [3] are used as the ideal models for the active parts of the amplifier. The following sections discuss the

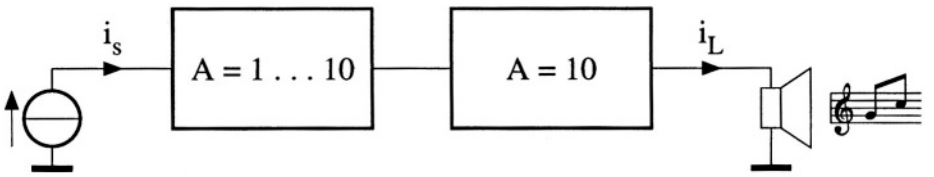

Figure B.1: The block diagram of the complete controllable output amplifier

implementation of the nullors. Section B.3 starts with some general comments on the implementation of the nullors. Section B.4 discusses the design of the output capability of the amplifier. In this section the focus is on the implementation of the class-AB operation for power efficiency. The class-AB operation is realized in the voltage domain and uses a version of the harmonic-mean relation. Subsequently, Section B.5 describes the behavior of the overall negative-feedback loop, i.e. loop gain, poles and stability. This is followed by the implementation of the biasing circuitry and some final implementation details in Section B.6. Section B.7 describes the physical realization and the measurement results of the amplifier. Finally, the conclusions are given in Section B.8.

B.2 The basic structure of the output section

In the LW receiver the output section has two functions:

- volume control
- driving the earphone.

The input current for the complete output amplifier is leveled by an AGC to a peak value of approximately 25 μA. A signal current of about 1 mA results in an acceptable sound level [4]. To have some margin, the maximal peak output current is chosen to be 2.5 mA, resulting in a required maximum amplification of 100. To be able to control the volume over a convenient range, the amplification is chosen to be controllable between 10 and 100.

The block diagram of the amplifier is shown in figure B.1. The first amplifier block has a variable gain between 1 and 10, whereas the second amplifier has a fixed gain of 10. As the signal levels in the first amplifier are still relatively low, the focus can be on the implementation of the variable gain instead of on the power consumption. For the second amplifier, the power consumption is the key item for optimization and this is not disturbed by the implementation of a variable gain. In this way the two functions, driving the load and controlling the gain, are realized in separate amplifiers and can, consequently be optimized independently.

Here is dealt with the fixed-gain amplifier. The specifications for the amplifier are:

- maximum input current ≈ 0.25 mA,
- source impedance 2 MΩ in parallel with 0.25 pF,
- Load = 30 Ω,
- Gain = 10,
- Bandwidth > 7 kHz,
- Distortion < 1 %,
- Supply voltage 1 to 1.5 V,
- Supply current as low as possible,
- Completely integrable in a bipolar process,
- Temperature range -10°C to +40 °C.

A straightforward interpretation of these specifications leads to the choice of a current amplifier. However, for a negative-feedback current amplifier, current sensing has to be done at the output. The output stage of the amplifier must be balanced, this can be done in the voltage or current domain [5], or the amplifier must have indirect feedback [6]. For the sake of power efficiency, it is favorable to realize the amplifier in a class-AB fashion.

Realizing a balanced or indirect class-AB output stage is a tedious job. Therefore, a different type of amplifier has to be used. The load impedance of the earphone is approximately 30 Ω and more or less constant over the frequency range of interest. Thus, the output of the amplifier may also be a voltage. In that case the feedback must sense the output voltage that is readily available. The final output current is determined by the impedance of the earphone. The resulting amplifier is a transimpedance amplifier.

For a maximum current of 2.5 mA the voltage across the earphone is only 75 mV. Grounding the earphone at one side results in saturation of one side of the AB output stage. Further, offset voltages at the output result in a relatively large offset current through the load as its impedance is only 30 Ω. Therefore, the amplifier has to be realized completely balanced, see figure B.2. As both amplifier halves are class AB, the signal current is comparable to that of the single-sided amplifier; only the quiescent current is doubled. Compared to the current amplifier, the voltage amplifier is more power efficient as it does not require an additional current path for the current feedback.

In figure B.2 the gain blocks are nullors [3]. A nullor is an ideal element which makes its input current and voltage zero by controlling its output current

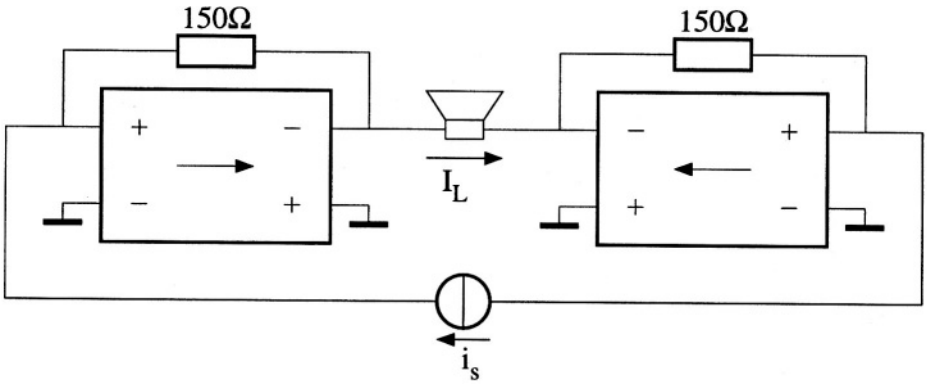

Figure B.2: The balanced transimpedance output amplifier

and voltage. In terms of gain parameters, the nullor has gain parameters which are infinite. Therefore, the gain from input current to output voltage of the balanced transimpedance amplifier is determined by the two feedback resistors only. To realize the required gain, the feedback resistors each have to be 150 Ω.

B.3 Implementation of the nullors

Now the basic structure of the amplifier has been chosen, the implementation of the nullors is that remains. The implementation of nullors can be done in several more or less independent steps [7]. As the signals at the input of the amplifier are already relatively large, optimization with respect to noise is not necessary. The remaining steps (condensed form) in chronological order are the design of:

- Output capability,
- Bandwidth and
- Biasing.

These items are discussed in the following sections.

B.4 Output capability

The output capability of the amplifier is determined by the maximum output signal that can be supplied. As the load impedance is only 30 Ω, the output capability is set by the maximum current which can be supplied. For this amplifier this must be about 2.5 mA. For power-efficiency purposes the output stage is chosen to be class AB. The current-gain factor of the transistors in the

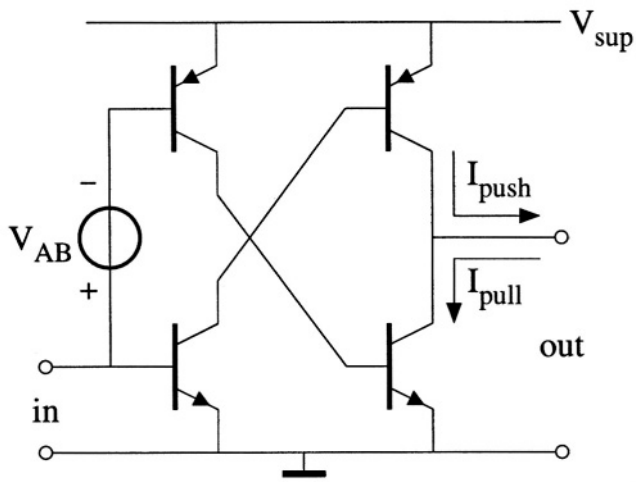

Figure B.3: The two last stages are chosen to be class-AB biased

DIMES01 process [8] are 80 for the PNP and 100 for the NPN. Consequently, the maximum input current of this stage amounts to 35 μA. When biasing the preceding stage in class-A mode, the bias current needs to be in the order of 50 μA. Therefore, this stage is also chosen to be class-AB. This is depicted in figure B.3. The voltage V_{AB} takes care of the class-AB operation [9]. The signs of the voltage source correspond to the situation in which the supply voltage is 1 V, as the sum of two base-emitters voltages is larger than 1 V. The implementation of this voltage source determines the final functioning of the AB control. When this voltage source is just a fixed voltage the classical AB control is obtained for which the following holds:

$$I_{push} * I_{pull} = \text{constant}. \tag{B.1}$$

Where I_{push} and I_{pull} are the currents flowing through the two output transistors. When large output swings are required, the minimum current can be very low, as the product of the two current is constant. Consequently, the f_T of the corresponding transistor may become too low, and an increase in distortion is found. When the amplifier is used in a feedback structure, even oscillations may occur due to the switch-on delay of this transistor. Therefore, very often the harmonic-mean relation is used [10]:

$$I_{push} * I_{pull} = (I_{push} + I_{pull}) * I_{min}. \tag{B.2}$$

When either I_{push} or I_{pull} becomes very large, the other current is limited to I_{min} preventing the corresponding transistor from becoming too slow.

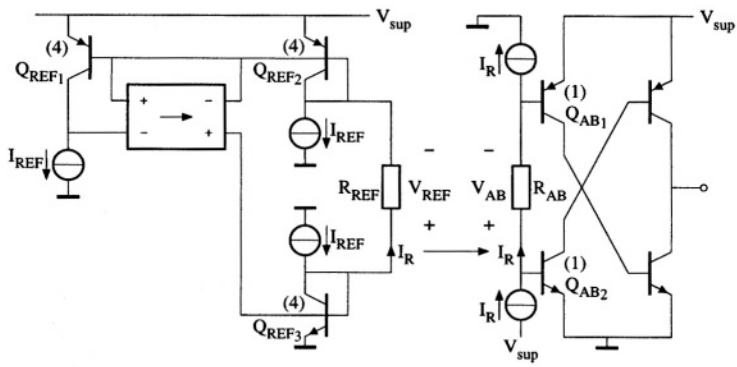

Figure B.4: The principle of the AB control voltage

B.4.1 The voltage source for the class-AB control

The implementation of the voltage source, which takes care of the AB control, determines the final functioning of the AB operation. The conventional class-AB stage uses two series-connected diodes [11] [1]. A reference current flowing through these two diodes determines the product of I_{push} and I_{pull}. Later on, other concepts were published that were able to work at supply voltages down to 1 V [9]. Here, however, the straightforward concept of [11] is extended to function for supply voltages down to 1 V. For these low supply voltages it is not possible to use series-connected junctions. The generation of the control voltage has to be done indirectly. The principle is depicted in figure B.4. The reference voltage obtained for the class-AB control is given by:

$$V_{REF} = V_{BE_{REF3}} + V_{EB_{REF2}} - V_{SUP} \tag{B.3}$$

with V_{SUP} the supply voltage. Both $V_{EB_{REF2}}$ and $V_{BE_{REF3}}$ are determined by I_{REF}. This is accomplished by transistor Q_{REF1} and the nullor. The nullor forces I_{REF} to flow completely through the collector of Q_{REF1}. Transistor Q_{REF2} is in parallel with transistor Q_{REF1} and thus its collector current also equals I_{REF}. The nullor supplies the current through the resistor, at both sides of the resistor, and thus the collector current of Q_{REF3} also has to be equal to I_{REF}. It must be noted that due to the low supply voltage, the voltage across the resistor is of opposite polarity with respect to conventional AB stages.

By copying the current through the resistor R_{REF} to the current through resistor R_{AB}, the following relation is found:

$$V_{SUP} - V_{EB_{AB1}} + V_{AB} - V_{BE_{AB2}} = 0 \tag{B.4}$$

[1] NB: traditionally two CC stages are used instead of two CE stages

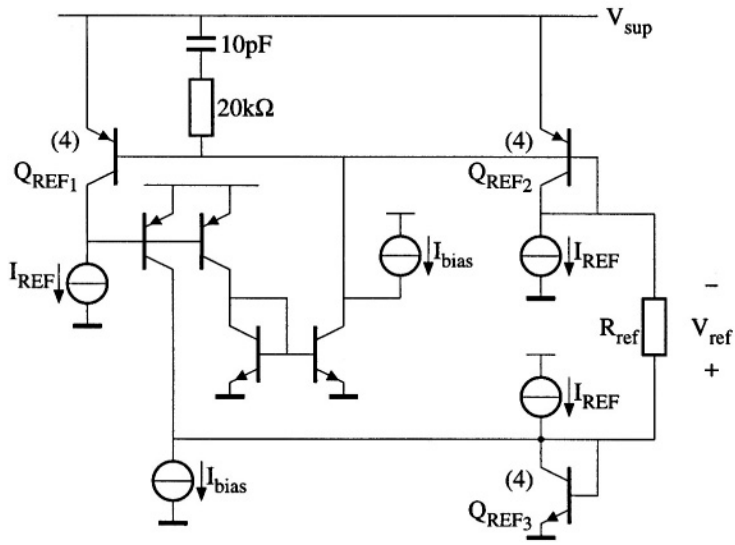

Figure B.5: The implementation of the nullor forcing I_{REF} through Q_1

or by substitution of the expression for V_{AB}:

$$V_{EB_{AB1}} + V_{BE_{AB2}} = V_{EB_{REF2}} + V_{BE_{REF3}}. \tag{B.5}$$

This is just the expression for AB control. The influence of the supply voltage, as it is in the AB control voltage, cancels.

In this design I_{REF} is chosen to be 1 μA. As a result the quiescent current of the first AB stage amounts to 0.25 μA, Q_{REF1-3} are four times as large as $Q_{AB1,2}$. The quiescent current of the second AB stage is approximately $25\mu A$; only 1% of the maximum output current. The resistor R_{REF} is chosen to be 100 kΩ. In that case the current through the resistor is in the order of the other currents in the reference source.

The implementation of the current copying from the reference source to the AB stage is discussed in paragraph B.4.3. The implementation of the nullor, which forces I_{REF} through the collector of Q_{REF1}, is depicted in figure B.5. In principle, the nullor can be implemented by one CE stage [6]. However, as both outputs of the nullor are required, an inverting and a non-inverting output terminal must be realized. The two outputs of the CE stage, the collector and emitter, are not usable due to the low supply voltage. Using a differential pair for implementing the nullor seems to be the next candidate [6]. However, in that case two base-emitter junctions are in series for a NPN stage, which is not possible for a 1 V supply, or the transistors are close to saturation in the case of

a PNP stage. The combination of two parallel CE stages and a current mirror for the inversion leads to a convenient solution.

The output stages of the nullor implementation have to be biased as they must be able to source and sink currents. For a 1.5 V supply the voltage across R_{REF} is approximately +0.2 V (recalling the polarity convention of V_{AB}). When the battery is almost empty, the supply voltage is reduced to 1 V and the voltage across R_{REF} is approximately -0.3 V. Therefore, the bias current is chosen to be about 5 μA to be able to cope with the complete range.

Frequency compensation of the loop comprising the nullor implementation and transistor Q_{REF1} is realized by pole splitting using pole-zero cancellation [7]. The closed loop exhibits a second-order Butterworth behavior with a bandwidth of approximately 1.4 MHz. The frequency compensation already took the influence of the current-copier implementations (to be discussed later on) into account.

B.4.2 The "harmonic-mean" control

As already mentioned, using strict AB control has the disadvantage of having transistors with very low biasing currents and thus becoming slow. For AB stages that are not fed back, this results in an increase in distortion. For this amplifier, which is fed back, the additional phase due to the switch-on delay caused oscillations. Therefore, some control analogous to the harmonic-mean relation has to be used. The product of the two transistor currents must not be constant but has to increase for increasing output current (which is approximately equal to the largest of both AB currents). The principal idea is depicted in figure B.6. In the figure only the first AB stage is depicted, Q_{AB1} and Q_{AB2}. The AB control voltage is modeled with a single voltage source and a resistor.

Two transconductance stages are placed in parallel to each AB transistor. Their transconductance is approximately equal to:

$$G_{stage} = \frac{1}{R_{fb} + 1/g_m} \tag{B.6}$$

where R_{fb} is the feedback resistor of the transconductance stage and g_m is the transconductance of the transistor. Assume the current through Q_{AB1} becomes relatively large; in that case the output current of its parallel-connected transconductance stages will also increase. The output current of the transconductance is fed through the output of the AB-control-voltage source, resulting in a change of its voltage such that the base-emitter voltage of Q_{AB2} does not decrease as much as it did originally. This results in a reduced decrease of its collector current.

The current must be sunk at one side of the resistor and sourced at the other side of the resistor. If it was to be sourced or sunk at one side only, the

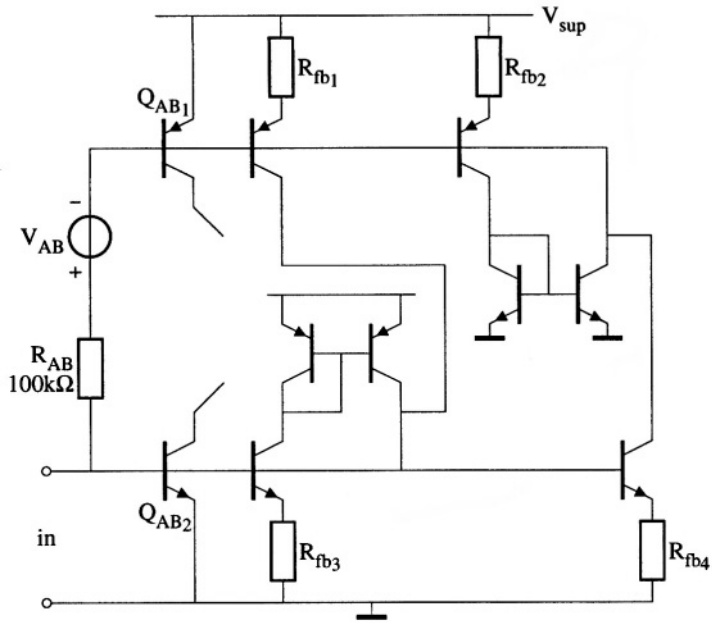

Figure B.6: Harmonic mean relation

input offset current of the complete AB stage, formerly only the base current, would change sign. Consequently, the overall loop gain of the amplifier would become positive when it was supposed to be negative. This sign reversal is signal dependent.

The harmonic-mean relation introduces a loop with a positive feedback. This can be seen as follows. When one of the two AB currents becomes large, the product of the two currents is increased by the harmonic-mean relation. This increase of the product can either cause an increase in the smallest AB current (this is the required option) or it can increase the current that is already large. Of course, a mix of both situations is also possible. When the large current becomes larger (the second option), the harmonic-mean relation increases the product even further and the current becomes even larger, and so on. This positive feedback loop must be counteracted by a stronger negative-feedback loop. This loop must keep the output current of the complete stage under control. Then the *smallest* AB current increases due to the harmonic-mean relation.

To be certain, the loop gain of the positive loop can be kept below one. The loop gain, *T,* of the positive loop shown in figure B.6 is approximately given by:

$$T = G_{stage} * R_{AB} \tag{B.7}$$

Now, with some straightforward reasoning, approximating expressions can be found for the behavior of this AB control. The quiescent current of the first AB stage remains $0.25\ I_{REF}$, when it is assumed for the moment that the transconductance stages do not influence the AB control when it is in its quiescent state. This assumption will be validated in a following paragraph. In the case of a relatively large signal excursion, the smallest current decreases the same factor as the largest current increases (for strict AB control) as the absolute variations in base-emitter voltages of the two AB transistors are equal. However, for this version of harmonic-mean control, the absolute variation of the smallest base-emitter voltage can be approximated by:

$$\Delta V_{BE_{small}} = (1 - T)\Delta V_{BE_{large}}. \tag{B.8}$$

The decrease in the smallest base-emitter voltage is reduced by $TV_{BE_{large}}$. The resulting collector current can then be written as:

$$I_{C_{small}} = (0.25 I_{REF})^{(2-T)}\, I_{C_{large}}^{(T-1)}. \tag{B.9}$$

For $T = 1$, the variation of the smallest base-emitter voltage is zero (equation B.8) as the variation due to the largest base-emitter voltage is completely compensated. Thus, $I_{C_{small}}$ remains $0.25\ I_{REF}$. In figure B.7 a sketch is shown of the collector current of $Q_{AB1,2}$ as a function of the input current, with the loop gain T as a parameter.

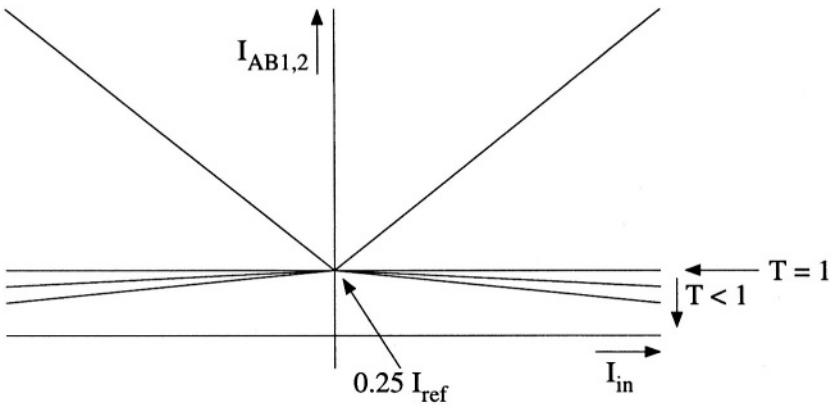

Figure B.7: The currents of the first AB-stage as a function of the input current, with T as a parameter, when the harmonic-mean control is used (only the asymptotes are depicted)

The expression for the loop gain of (B.7) holds when base currents and so on are ignored. However, due to the asymmetric drive of the amplifier, the base current of Q_{AB1} flows through the output of the AB-control voltage source. This current is in the opposite phase with respect to the current for the harmonic-mean control from the transconductance stages reducing their influence. In order to counteract this, the resistors in the transconductance stages at the PNP side are chosen to be 50 kΩ; at the NPN side these resistors are chosen to be 80 kΩ. These resistors are somewhat smaller than 100 kΩ as the g_m of the transistor reduces the complete transconductance, see (B.6).

B.4.3 The current copiers

Now that the AB control has been implemented, attention can be paid to the implementation of the current copiers. In figure B.8 the AB-control voltage source is depicted again but now the current copiers are also drawn. The collector currents of transistor Q_{m1} and Q_{m2} minus the current from their bias sources equals the current flowing through R_{REF}. These copied currents flow through a resistor equal to R_{REF} and thus the voltage across R_{REF} is copied and this copied voltage is allowed to float. In principle, the current flowing in or out at the other side of the resistor has to be copied from Q_A. However, when implementing the rest of the output amplifier, this current source is in parallel with the current source for the biasing of the first amplifying stage (see figure B.9, transistor Q_1 and Q_2), and does not need to be implemented explicitly.

Transistors Q_{x1} and Q_{x2} are required for the following reason. When the AB amplifier is in its quiescent state, some current already flows through the four transconductance stages (see figure B.6). These currents result in an ad-

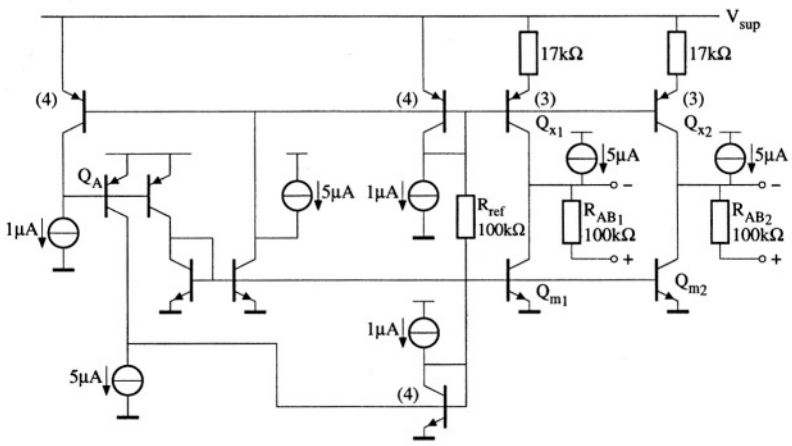

Figure B.8: The realization of the current copiers

ditional voltage drop across R_{AB} of the AB-control voltage source, resulting in a quiescent current larger than $0.25\ I_{REF}$. This additional voltage drop can be taken into account when calculating the quiescent current. However, then the relation between the quiescent current and I_{REF} is not very convenient. It is better to compensate for this additional voltage drop. The collector current of Q_{x1} and Q_{x2} are approximately equal to the sum of the *quiescent* currents flowing through the NPN and the PNP transconductance stages, a scaling of three and a resistor of $17\ \mathrm{k}\Omega$ resulted in the desired behavior. This current is used to cancel the quiescent current of the transconductance stages through R_{AB}. Now the quiescent current of the first AB stage is equal to I_{REF} again, except for the scaling factor of 0.25, of course.

B.5 The overall-loop behavior

The overall-loop behavior is determined by its dc loop gain, poles and zeros. When implementing the amplifier with only the two AB stages, the dc loop gain $A\beta$ is very low. It is approximately:

$$A\beta = \beta_{NPN} \cdot \beta_{PNP} \cdot \frac{0.5R_L}{r_\pi} \tag{B.10}$$

where r_π is the input resistance of either the NPN or PNP input stage, which depends on the polarity of the input signal. When $\beta_{NPN} = 100$, $\beta_{PNP} = 80$, $0.5R_L = 15\Omega$ and $r_\pi = 8M\Omega$, the loop gain is approximately 0.01. This is, of course, not practical and an additional amplifying stage has to be added. This

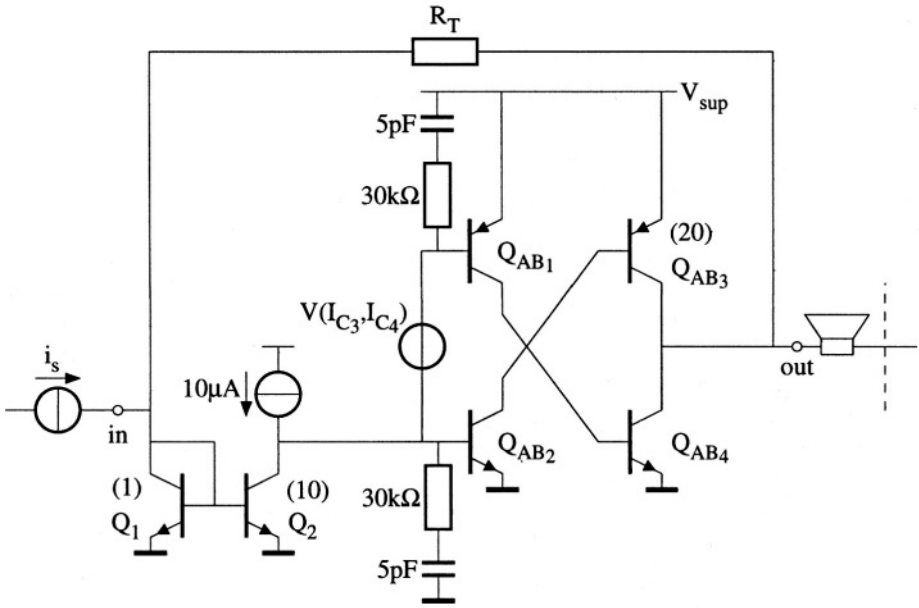

Figure B.9: The implementation of the input stage to obtain acceptable loop gain

stage adds gain to the loop as it is amplifying. Further, by choosing its bias current to be larger than $0.25\ \mu A$, as it is for the (quiescent) first AB stage, its input impedance reduces. Consequently, the fraction in expression (B.10) also increases.

A convenient choice is a current mirror with a scaling factor of ten and a bias current of the input transistor equal to $1\ \mu A$. One half of the amplifier is depicted in figure B.9. For the sake of clarity, the harmonic-mean control is omitted in the figure. The loop gain increases due to the gain of the current mirror by a factor of ten. Due to the lowered input impedance of the input stage, the loop gain increases by an additional factor of about four hundred. Due to the output impedance of the current source at the output of the first stage the loop gain is reduced by a factor of 4. The resulting loop gain is about 10. Frequency compensation is obtained by the pole-zero cancellation networks at the inputs of Q_{AB1} and Q_{AB2}. Due to the varying bias current of the AB stages, the bandwidth of the amplifier varies as a function of the signal. At zero signal the bandwidth is approximately 100 kHz, i.e. the poles are at 80 kHz ± 60j kHz. For a large output signal (1.5 mA peak), the bandwidth is 2.2 MHz and poles are found at 1 MHz ± 2j MHz.

B.6 The complete circuit

The complete circuit consists of two amplifier halves, one of which is depicted in figure B.9. To be able to drive a maximum current of about 2.5 mA to the load, the PNP output transistor (Q_{AB3}) is chosen to be twenty times as large as the smallest PNP transistor in the DIMES01 technology (in order to prevent it from high-level injection at maximum output current). The maximum output current is limited by the saturation of the NPN output transistor (Q_{AB4}). The collector voltage of this transistor is more or less set by the base-emitter voltage of the input transistor, Q_1, via the low-ohmic feedback resistor. This voltage is about 0.58 V. When sinking an output current of several mA, the base-emitter voltage of Q_{AB4} becomes relatively large and the base-collector junction starts conducting. Q_{AB3} does not have this problem as its saturation current (I_S) is about 200 times as large as the saturation current of Q_{AB4}. Thus Q_{AB3} saturates before Q_{AB4} saturates.

The quiescent current of the first AB stage is approximately $0.2\ \mu\text{A}$, which is slightly lower than the intended $0.25\ \mu\text{A}$, due to the influence of base-currents. The currents are bounded at the lower side to about $0.05\ \mu\text{A}$. For ideal transconductance stages, the lower bound would have been $0.2\ \mu\text{A}$ as the loop gain was designed to be 1. However, due to the signal dependency of the transconductance, see equation B.6, the loop gain is for the relatively small signals smaller than 1. For this application this poses no problems.

The biasing sources, depicted in the previous figures, are all derived from one reference current by means of current copiers. For the measurements this current was supplied by an external source. Later on, when the complete receiver will be integrated, a master reference source will be made.

B.7 Measurement results

In figure B.10 a chip photo is depicted of the amplifier. The two large parts are the two amplifying halves and the smaller part is the circuit generating the AB-control voltage. When measurements were performed, the amplifier exhibited common-mode instability. The reason for this is that when the circuit is performing normally, the load for each amplifier half is about $15\ \Omega$. However, when one half does not function properly, due to startup, for instance, its output impedance is not low and as a consequence, the load for the other amplifier half becomes very large. The loop gain of the amplifier half increases by a factor of 1000 due to this increase in the load impedance. When the complete amplifier functions properly, the low load impedance reduces the loop gain by a factor of 1000. For normal operation, the amplifier was frequency-compensated to a second-order behavior. However, due to the increase of loop gain, for the common-mode case, a third pole becomes important, driving the other two poles

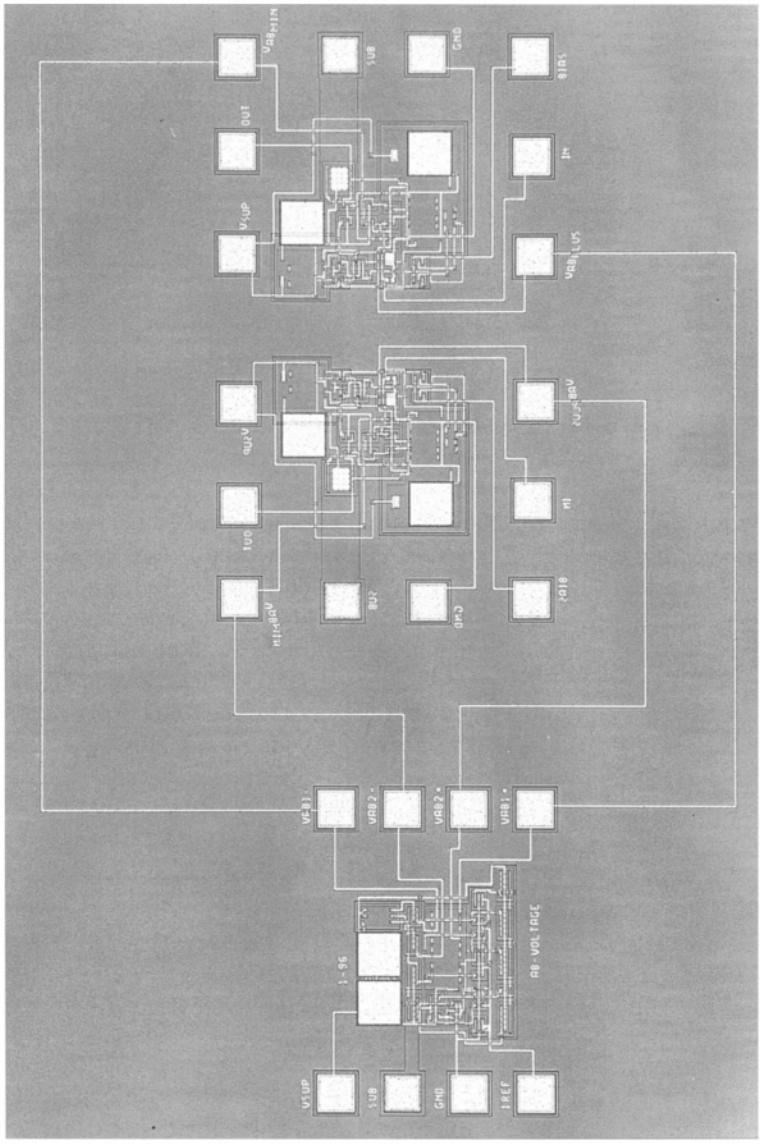

Figure B.10: Chip photo of the complete amplifier

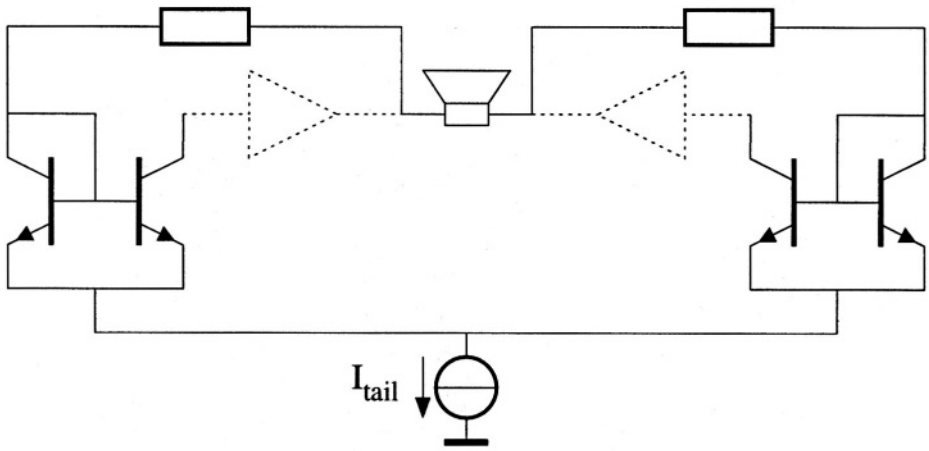

Figure B.11: Solution to the common-mode instability

into the right half plane. For the measurements this problem was counteracted by connecting the middle of the load, via a large capacitor, to ground. The load impedance for an amplifiers half is then determined independently of the other amplifier half. This large capacitor, however, cannot be used in the final integrated LW receiver. For that, a redesign has to be done using the principle as depicted in figure B.11. The two input stages, each comprising Q_{1_i} and Q_{2_i} (see figure B.9), are combined into one differential stage. The dotted drawn amplifiers represent the AB stages. For common-mode signals the input impedance of this stage is determined by the impedance of the tail-current source which can be rather high. For an Early voltage of 50 V, the impedance of both input terminals to ground is about 50 $M\Omega$ as only 1/22 part of the leakage current through the impedance of the current source is seen (I_{tail} =22 μA). At the output of the amplifier a relatively small capacitor will introduce a relatively low-frequency pole, reducing the common-mode bandwidth. However, the input impedance to ground has a pole at a frequency equal to the frequency of the pole in the impedance of the current source. Thus, beyond this frequency the common-mode loop gain increases (a zero). To be able to counteract the increase of loop gain found by a factor of thousand, the pole at the output of the amplifier has to be a factor of a thousand lower than the pole in the current-source impedance. This can be done with relatively small capacitances (a few pF). Of course, this low-frequency pole is not seen in the differential mode loop. The additional capacitor at the output only loads the much lower differential mode impedance, shifting the corresponding pole a bit downwards, making a reexamination of the frequency compensation necessary, of course.

In figure B.12 the spectrum of the output signal is depicted for an output

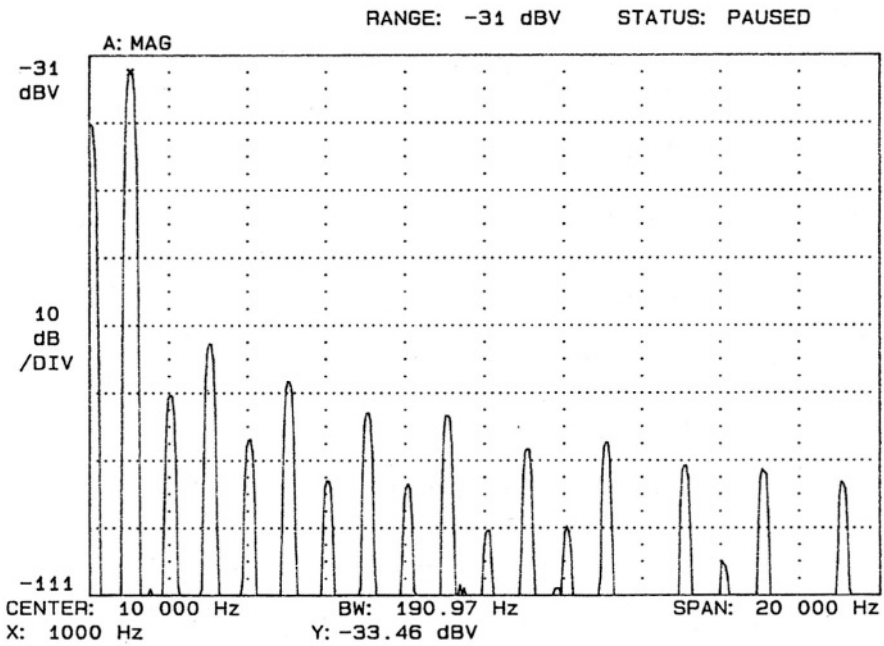

Figure B.12: The spectrum of a 1 mA output current at 1 kHz

signal at 1 kHz with an amplitude of 1 mA. The total harmonic distortion (using the harmonics up to 20 kHz), remains below to 0.9 %. This figure is more or less independent of the supply voltage. This figure was measured for a supply voltage of 1 V. Increasing the supply voltage to 1.5 V resulted in an improvement of some tenths of a dB.

Figure B.13 depicts the transfer as a function of the frequency with the amplitude of the input signal as a parameter. The amplitude of the input signal was logarithmically varied from 3.5 μA to 20 μA in steps of 4 dB. The bandwidth varied from 100 kHz to 300 kHz. The bandwidth was lower than expected due to the larger parasitics. The frequency compensation, however, proved to be good enough.

Finally, in table B.1 an overview of the measurement results is given.

B.8 Conclusions and discussion

In this appendix the design of a low-voltage low-power negative-feedback class-AB amplifier is described. The supply voltage can be as low as 0.9 V. The

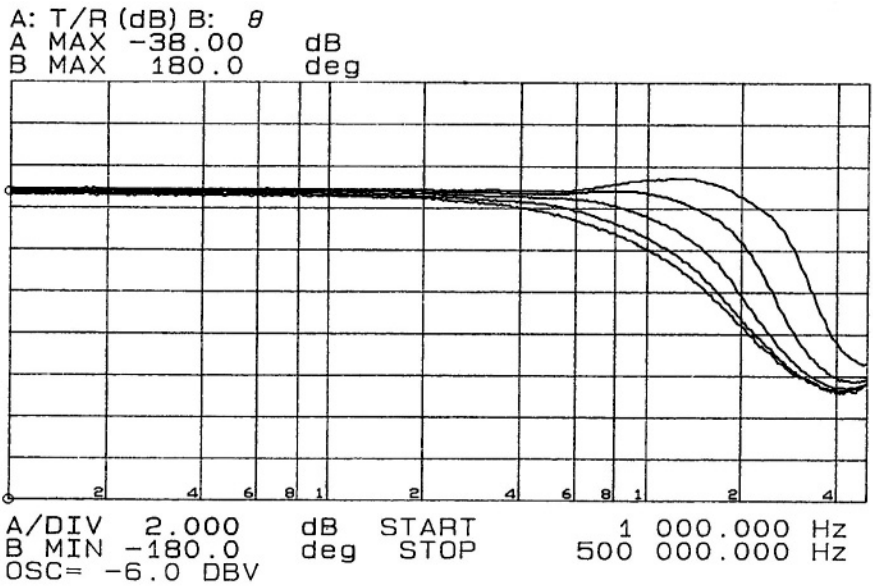

Figure B.13: The transfer as a function of the frequency with the amplitude of the input signal as a parameter

Supply voltage	0.9 V - 1.5 V
Total quiescent current	$\approx 100\mu$ A
Maximum output current	> 2.5 mA
THD @ 1 kHz and I_{out}=1 mA, $V_{sup} = 1$ V	0.9 %
Bandwidth I_{out} =35 μA	$\approx$ 100 kHz
Bandwidth I_{out} =200 μA	$\approx$ 300 kHz
Chip size	3 mm^2

Table B.1: Performance of the amplifier

quiescent current of the amplifier is only 100 μA, whereas the maximum output current is greater than 2.5 mA. The amplifier is capable of driving an earphone with an impedance of 30 Ω. The quiescent current of the output stage is only 1% of the maximum output current.

This low quiescent current, relative to the maximum output current, is obtained by using class-AB biasing for the last two stages of the total of three stages. To prevent the transistors from becoming too slow, which is inherent in strict class-AB operation, a new type of implementation of the harmonic-mean relation is described.

Due to the low-voltage constraint, the circuit for controlling the AB voltage, in order to obtain the harmonic-mean relation, is realized in an indirect way. This is necessary since it is not possible to stack conducting junctions between the supply rails.

The chip showed a common-mode instability, which for this chip was solved by grounding the middle of the load with a relatively large capacitor. For the final LW receiver, a balanced input stage is discussed which protects the amplifier from common-mode instability.

Bibliography

[1] A. van Staveren, G.L.E. Monna, C.J.M. Verhoeven, and A.H.M. van Roermund. A low-power class-ab negative feedback amplifier for a 1V LW receiver. *Analog Integrated Circuits and Signal Processing,* 20:63–75, 1999.

[2] W.A. Serdijn. *The Design of Low-Voltage Low-Power Analog Integrated Circuits and Their Applications in Hearing Instruments.* PhD thesis, Delft University of Technology, February 1994.

[3] H.J. Carlin. Singular network elements. *IEEE Transactions on Circuit Theory,* 11:67–72, March 1964.

[4] M.P. Lubbers. An output amplifier for a 1 V portable AM receiver. Master's thesis, Delft University of Technology, December 1994. In Dutch.

[5] A.C. van der Woerd and A.C. Pluygers. Biasing a differential pair in low-voltage analog circuits: A systematic approach. *Analog Integrated Circuits and Signal Processing,* 3:119–125, 1993.

[6] E.H. Nordholt. *The Design of High-Performance Negative-Feedback Amplifiers.* Elsevier, Amsterdam, 1983.

[7] C.J.M. Verhoeven, A. van Staveren, and G.L.E. Monna. Structured electronic design, negative-feedback amplifiers. Lecture notes ET4 041, Delft University of Technology, 1999. To appear at John Wiley & Sons LTD, Chichester.

[8] L.K. Nanver, E.J.G. Goudena, and H.W. van Zeijl. DIMES-01, a baseline BIFET process for smart sensor experimentation. *Sensors and Actuators, Part A, Physical,* 36(2):139–149, April 1993.

[9] W.C.M. Renirie, K.J. de Langen, and J.H. Huijsing. Parallel feedforward class-AB control circuits for low-voltage low-power rail-to-rail output stages of operational amplifiers. *Analog Integrated Circuits and Signal Processing,* 8:37–48, 1995.

[10] E.Seevinck, W.de Jager, and P.Buitendijk. A low-distortion output stage with improved stability for monolithic power amplifiers. *IEEE Journal of Solid-State Circuits,* 23:794–801, June 1988.

[11] J.E. Solomon. The monolithic op amp: a tutorial study. *IEEE Journal of Solid-State Circuits,* 9:314–332, December 1974.

Appendix C

The Effective Q versus the phase shift

In this appendix the effective Q of a resonator is calculated as a function of its detuning phase. The starting point is the impedance of the intrinsic resonator:

$$Z_s = sL_s + \frac{1}{sC_s} + R_s, \tag{C.1}$$

where s is the Laplace variable, L_s, C_s and R_s, are the inductor, capacitor and resistor of the series resonator, respectively. The phase shift, ϕ, as a function of the frequency, is given by:

$$\phi = \arctan\left(\frac{\omega R_s C_s}{1-\omega^2 L_s C_s}\right) - \frac{\pi}{2}. \tag{C.2}$$

The quality factor as a function of the frequency is defined as:

$$Q(\omega) = \frac{1}{2} \cdot \omega \cdot \frac{d\phi}{d\omega}. \tag{C.3}$$

Applying this to equation (C.2) yields:

$$Q_s(\omega) = \frac{1}{2} \frac{\left[\frac{\omega R_s C_s}{1-\omega^2 L_s C_s} + \frac{2\omega^3 R_s C_s^2 L_s}{(1-\omega^2 L_s C_s)^2}\right]}{1+\left(\frac{\omega R_s C_s}{1-\omega^2 L_s C_s}\right)^2}. \tag{C.4}$$

From equation (C.2) it follows that:

$$\tan\left(\phi + \frac{\pi}{2}\right) = \frac{\omega R_s C_s}{1-\omega^2 L_s C_s} \triangleq \tan(\alpha). \tag{C.5}$$

Applying this in equation (C.4), yields the following for the effective quality factor:

$$Q_{eff} = \frac{1}{2}\cos^2(\alpha) \cdot \left[\tan(\alpha) + 2 \cdot \frac{\omega L_s}{R_s} \cdot \tan(\alpha)\right]. \tag{C.6}$$

From equations (5.45) and (5.49), the following expression can be found for the frequency of oscillation, ω_{osc}:

$$\omega_{osc} = \frac{\omega_0}{\sqrt{1 - \frac{\tan(\alpha - \pi/2)}{Q_s}}}. \tag{C.7}$$

Substitution of this expression in equation (C.6), yields:

$$Q_{eff} = \frac{1}{2}\cos^2(\alpha) \cdot \left[\tan(\alpha) + 2 \cdot \frac{\omega_0 L_s}{R_s\sqrt{1 - \frac{\tan(\alpha - \pi/2)}{Q_s}}} \cdot \cdot \tan(\alpha)\right]. \tag{C.8}$$

Now assuming that $0 < \phi \ll \frac{\pi}{4}$ or $0 < \alpha - \frac{\pi}{2} \ll \frac{\pi}{4}$, yields:

$$Q_{eff} = \cos^2(\phi) \cdot Q_s. \tag{C.9}$$

Appendix D

Design example: second-order compensated BGR

D.1 Introduction

In this appendix a design example of a second-order compensated bandgap reference [1] is described. Key issue for this design example is the compensation of the temperature behaviour by means of a linear combination of base-emitter voltages.

The basic structure used for the second-order compensated bandgap reference is given in Figure D.1. The design objective is a reference voltage of 200 mV. From temperature compensation the scaling factors can be found to be [equations (6.40) and (6.41)]:

- $a_1 = 0.81$

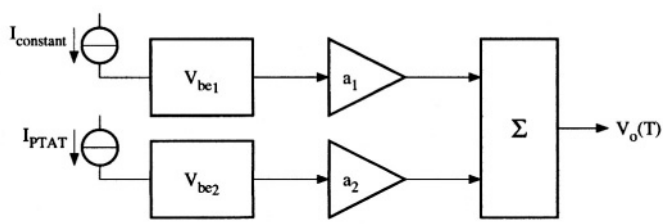

Figure D.1: A linear combination of two base-emitter voltages, a_1 and a_2 are the two scaling factors

- $a_2 = 0.64$

For this design the following design parameters were used:

- $I_{S1} = I_{S2} = 18aA$
- $\eta = 3$
- $I_{C1} = 15\mu A$
- $I_{C2} = 0.2\mu A$

In the subsequent sections realistic implementations of the separate building blocks of figure D.1 are given. At the interface between two blocks, special attention is paid to the possible interaction between the two blocks.

D.2 The design of the V_{BE} generator

In chapter 6 is was shown that the relation between the base-emitter voltage and the collector current is important. Therefore, the collector current has to be biased accurately. A bias circuit is needed that makes the collector current equal to the desired value.

The base current of a transistor has a temperature dependency that is different than that of the collector current. Therefore, the base current is not allowed to have effect on the collector current. The bias circuit has to supply the base current.

Further, the load current of the base-emitter voltage generator must not influence the collector bias current. The bias circuit has to supply the load current too.

Finally, to be able to ignore the forward Early effect, the bias circuit of the base-emitter voltage generator has to make $V_{bc} = 0$.

The biasing of the transistor is depicted in Figure D.2. A nullor is used for the biasing circuit. A nullor is a two-port that regulates the input voltage and current to zero by regulating the output voltage and current (see also [2]). The input current of the nullor is zero so the bias current I_C flows completely through the collector. The base and the load current are supplied by the nullor. Further, the input voltage of the nullor is zero, resulting in a zero base-collector voltage.

The next step in the design is to implement the nullor with a circuit. See Figure D.3.

For the input stage of the nullor implementation (a differential pair), MOS transistors are preferable because of the absence of input bias currents. However, the available process was a bipolar process so bipolar transistors had to be used here.

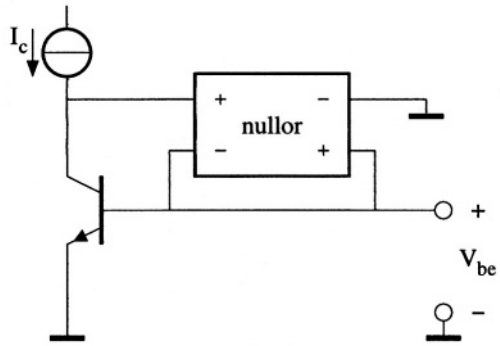

Figure D.2: The base-emitter voltage generator with a nullor for the biasing

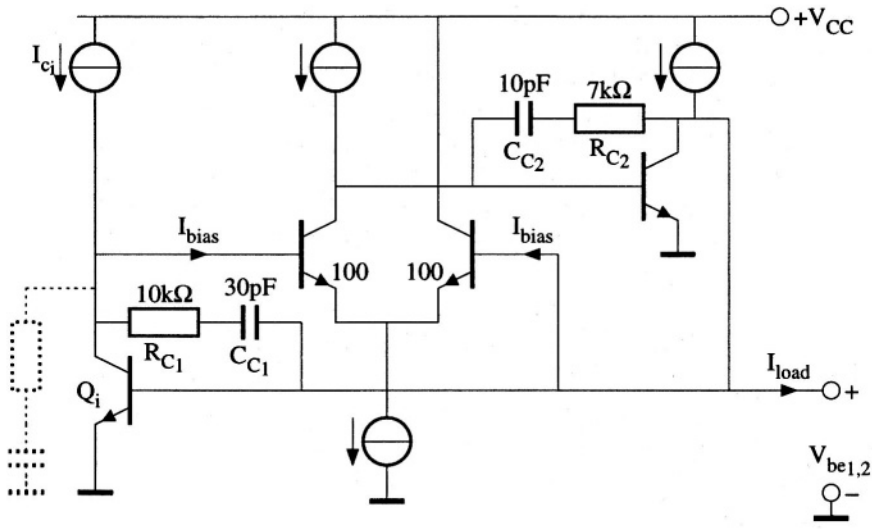

Figure D.3: The base-emitter voltage generator with a circuit implementation for the nullor

To minimize the influence of the input bias current I_{bias}, and the load current I_{load}, a second amplifying stage (a CE-stage) is used. To prevent the differential pair from saturating at lower temperatures, an NPN is chosen for this second stage.

To determine the optimum values of the bias currents, the equivalent noise voltage at the output of the base-emitter voltage generator is examined. The equivalent noise power density spectrum $S_{V_{be_i}}$ (in V^2/Hz) at the output of the V_{be} generator is approximately:

$$S_{V_{be_i}} = 2kTr_{e_i} \tag{D.1}$$

with $(r_{e_i} = kT/qI_{C_i})$ and I_{C_i} the collector bias current of Q_i, and it is independent of the noise contribution of the other transistors. Hence the bias current of the input stage of the nullor implementation can be chosen to be very small in order to obtain negligible input bias currents. Expression (D.1) assumes that low-noise current sources are available.

The value of the bias current of the output stage is based on the load and the base current of Q_i.

The available voltage for the tail current source of the differential pair is very small. To obtain some extra voltage, the emitters of the transistors of the differential pair are enlarged. For each time the emitter area increases by a factor 10, 60mV is obtained. When MOS transistors with low threshold voltages are available, the available voltage for the tail current source can be made large enough without the need to enlarge transistors.

High-frequency stability is obtained by the pole-splitting networks R_{C1}, C_{C1} and R_{C2}, C_{C2} for the circuit generating V_{be1}. Because Q_i in the circuit generating V_{be2} is biased at a different current than Q_i in the corresponding circuit for V_{be1} (resulting in another pole-zero pattern), the pole-splitting network R_{C1}, C_{C1} has to be replaced by a pole-zero cancellation (the dotted network in Figure D.3).

D.3 The design of the combiner

The combiner scales and adds the two base-emitter voltages. The scaling of the base-emitter voltages is realized passively. This passive scaling is implemented as a resistive divider and is shown in Figure D.4.

The scale factor a_i equals:

$$a_i = \frac{R_{i1}}{R_{i1} + R_{i2}} \tag{D.2}$$

The nullor is realized by a three-stage circuit for obtaining high loop gain. See Figure D.5. The high loop gain is necessary for reducing the non-linear

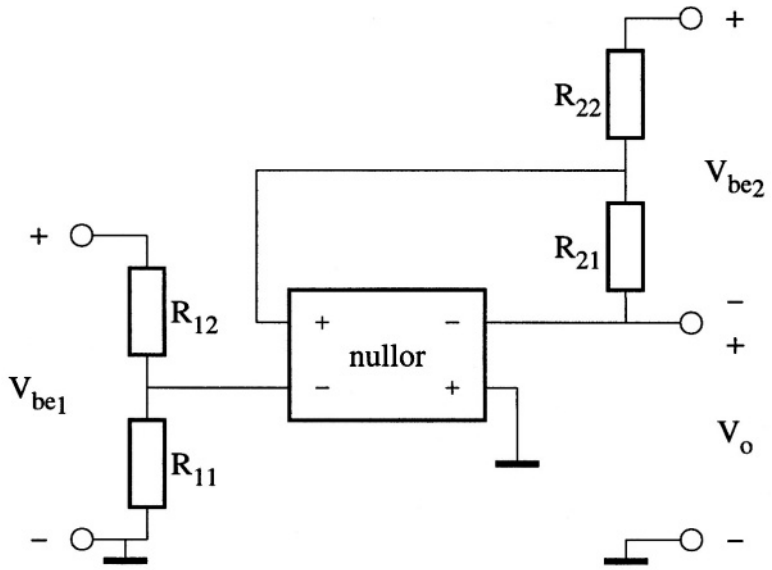

Figure D.4: The combiner for the scaling and addition of the two $V_{be}s$

offset voltage of the differential pair. This offset voltage is in series with the two scaled base-emitter voltages and is caused by the current from the generator for V_{be2}, flowing through the output stage of the combiner. To ease frequency compensation for the third stage a current mirror with a scaling factor of 10 is chosen.

Because the divider is loaded now with an input bias current I_B of the nullor implementation (Figure D.5) a voltage equal to

$$V_x = I_B(R_{21}//R_{22} - R_{11}//R_{12}) \tag{D.3}$$

is added to the output voltage V_O.

For the biasing currents the noise behavior is examined. All relevant noise sources (Figure D.6) are transformed into an equivalent noise source $V_{n,eq}$ at the output with a power spectrum (V^2/Hz):

$$\begin{aligned} S_{V_{n,eq}} &= S_{R_{11},R_{12}} + S_{R_{21},R_{22}} + S_{Q_A,Q_B} + S_{V_{be_1}} + S_{V_{be_2}} \\ &= 4kT[R_{11}//R_{12} + R_{21}//R_{22} + 0.5(r_{e_A} + r_{e_B} + a_1^2 r_{e_1} + a_2^2 r_{e_2})] \end{aligned} \tag{D.4}$$

with $r_{e_{A,B}} = kT/qI_{C_{A,B}}$ of the input transistors Q_A and Q_B.

To minimize the noise power at the output, each part of the combiner circuit should contribute less than the two base-emitter voltage generators contribute. This means that for the resistors of the dividers should hold:

$$R_{11}//R_{12} + R_{21}//R_{22} < 0.5(a_1^2 r_{e_1} + a_2^2 r_{e_2}) \tag{D.5}$$

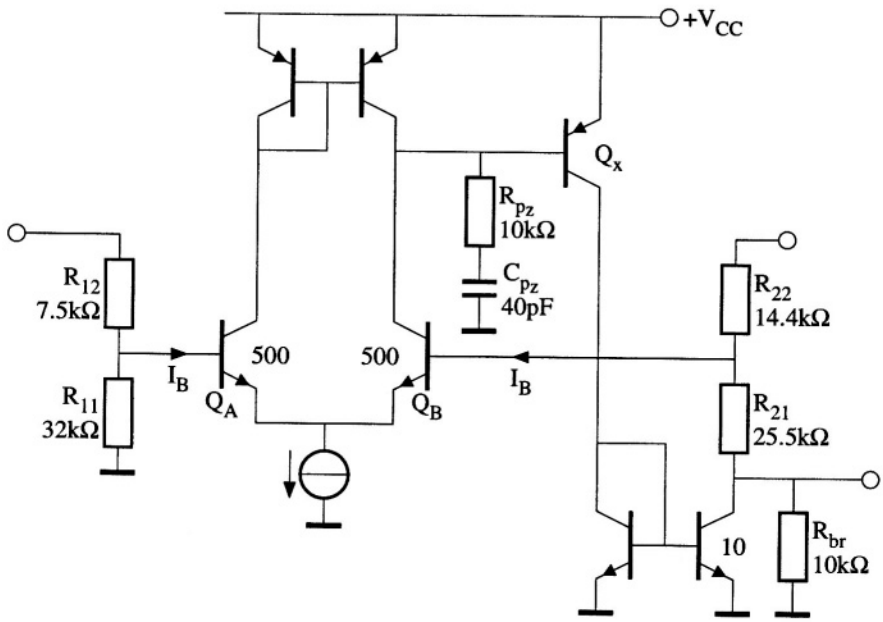

Figure D.5: The combiner with a circuit implementation for the nullor

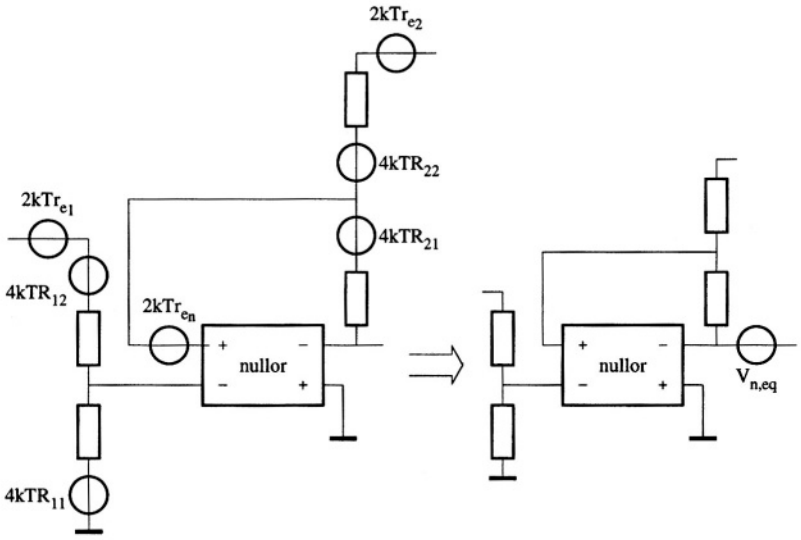

Figure D.6: The relevant noise sources of the bandgap reference

and for the bias currents I_{CQ_A,Q_B} of the differential pair:

$$I_{CQ_A,Q_B} > 2\left(\frac{I_{C_1}I_{C_2}}{a_1^2 I_{C_2} + a_2^2 I_{C_1}}\right) \tag{D.6}$$

The output of the combiner is the output of the bandgap reference. Therefore the bandwidth of the output impedance of the combiner has to be as large as possible. If this is the case, only a small capacitor in parallel with the output is needed to obtain a low output impedance for the frequencies beyond that bandwidth. High-frequency stability is obtained by pole-zero cancellation (R_{pz}, C_{pz}) and by resistive broadbanding (R_{br}). Although the resistive broadbanding reduces the loop gain, this reduction of loop gain has no effect on the offset voltage of the differential pair because the resistive broadbanding is placed at the node where the current from the V_{be2} generator is injected. The part of the current causing the offset voltage is decreased by the same factor as by which the loop gain is reduced.

D.4 Design of the bias circuits

One of the two base-emitter voltage generators has to be biased with a current proportional to the absolute temperature, PTAT $(\theta = 1)$, the other has to be biased with a constant current $(\theta = 0)$ (see for instance section 6.3.3.4). So, essentially, two types of bias currents have to be generated. All the other bias currents can be derived from these two current sources.

D.4.1 The constant current source

The bias current with $\theta = 0$ is easily derived from the output voltage via a transadmittance amplifier. However, this introduces a loop (Figure D.7).

To see if there is a unique DC solution for this loop the output voltage as a function of the current I is calculated. The output voltage is given by:

$$V_O = a - b\ln\left(\frac{I_C}{I_S}\right) \tag{D.7}$$

with $a > 0$, $b > 0$ and $I_C = -I$. Further the current I is given by

$$I = \frac{V_O}{R_t} \tag{D.8}$$

with R_t the feedback resistor of the transadmittance amplifier. The graphically determined solution of these two equations is shown in Figure D.8. It can be seen that there is only one DC solution.

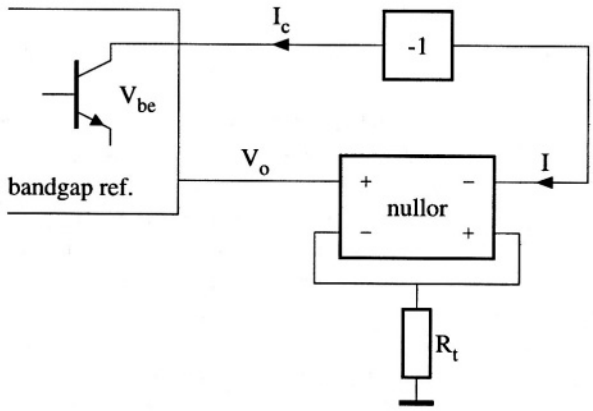

Figure D.7: Bias loop for the biasing of the (V_{be}) transistor with a constant collector current ($\theta = 0$)

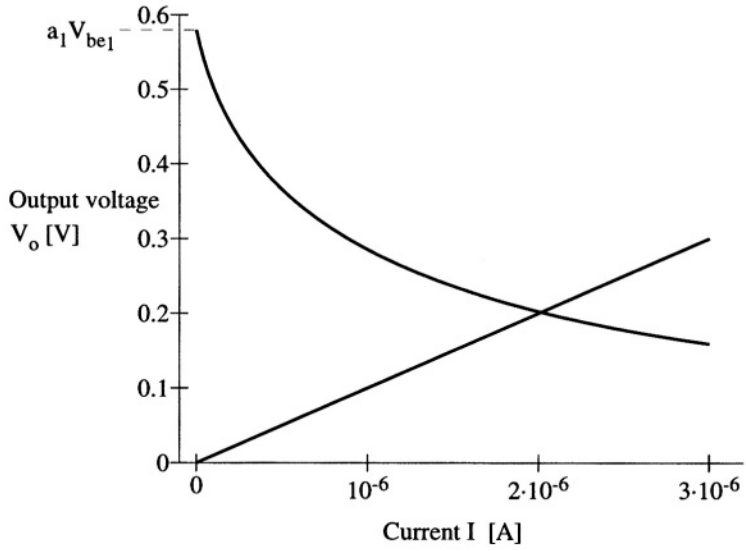

Figure D.8: The output voltage V_O versus the current I

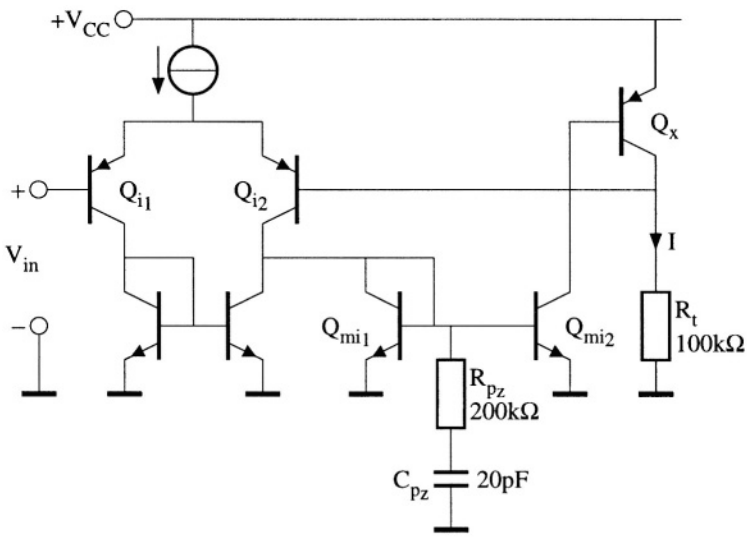

Figure D.9: The transadmittance amplifier

In Figure D.9 an implementation of the transadmittance amplifier, shown in Figure D.7, is depicted. Again, the nullor is implemented by a three-stage amplifier. The current mirror (Q_{mi1} and Q_{mi2}) is necessary to prevent the differential pair (Q_{i1} and Q_{i2}) from going into saturation. In the case of a direct connection between the differential pair and the output transistor Q_x (in which case the other output of the differential pair has to be used to keep the loop gain negative), the base-collector voltage of Q_{i1} would decrease at higher temperatures due the constant base voltage and the decreasing base-emitter voltage of Q_x, which would result in saturation. High frequency stability is obtained by means of pole-zero cancellation through R_{pz} and C_{pz}.

D.4.2 The PTAT current source

The current source with $\theta = 1$ is a PTAT current source. The PTAT current source shown in Figure D.10 [3] is used. This current source needs a small current I_{start} to prevent start-up problems.

In this circuit the current I_{PTAT} is given by:

$$I_{PTAT} = \ln(2)\frac{kT_r}{qR_{PTAT}}\left(\frac{T}{T_r}\right) \tag{D.9}$$

To obtain high-frequency stability pole-splitting is used (C_{sp} and R_{sp})

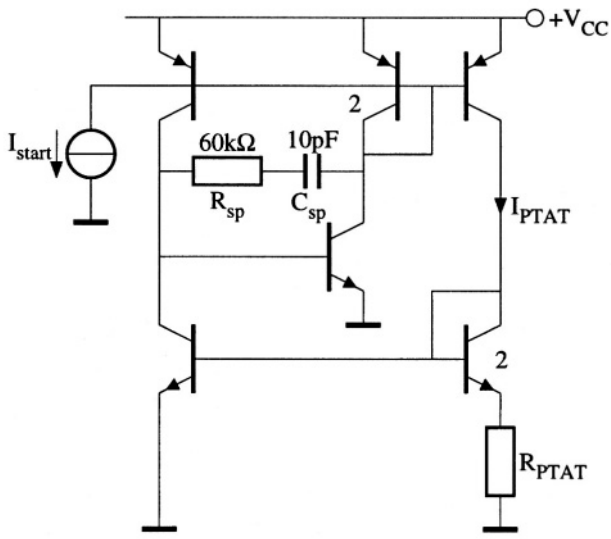

Figure D.10: A PTAT current source

D.5 The total circuit

The combination of all the sub-circuits yields the circuit shown in Figure D.11. The start-up current for the PTAT is realized by means of two diodes Q_{start1}, Q_{start2} and a resistor R_{start}. This start-up circuit is only applicable in 1V circuits. For higher supply voltages, more diodes have to be taken in series. In designing the current sources the temperature dependencies of R_{PTAT} and R_T were not taken into account. When these dependencies are taken into account the scaling factors a_1 and a_2 slightly change. For exact calculations the system of equations has to be changed, but this is beyond the scope of this appendix.

D.6 Realization and measurement results

D.6.1 Adjustment of the circuit

Ideally, no errors other than the higher-order temperature behavior of the base-emitter voltages occur (third and higher). In the previous sections all the ideal building blocks were implemented by practical circuits, introducing matching errors (resistor ratios for a_1 and a_2) and absolute errors (device parameters and the resistor in the PTAT source for example). Hence, the realized circuit should be adjusted for optimum circuit performance, i.e. for a temperature-independent output voltage up to the second order.

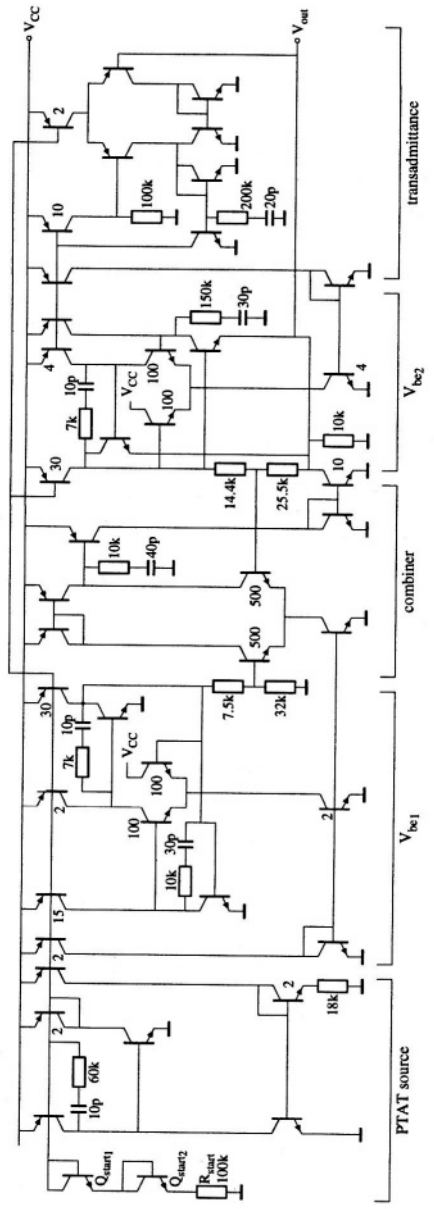

Figure D.11: The total circuit of the second-order compensated bandgap reference

For the adjustment of the bandgap reference one resistor needs to be trimmed. This can be the $18k\Omega$ resistor in the PTAT source or the $100k\Omega$ resistor in the transadmittance amplifier. These resistors determine the collector bias current of the reference transistors and by that the constant and the first-order term of the base-emitter voltage. With this adjustment the first-order temperature behavior can be minimized. A constant error may remain after this adjustment.

In the case of too large matching errors between the resistors implementing the ratios a_1 and a_2 or too large absolute errors in η or α_n, an additional adjustment for minimizing the second-order behavior is necessary. This adjustment can be done by trimming one of the resistors of a_1 and a_2, and has to be done before the first-order behavior is tuned because it affects the first-order behavior.

D.6.2 Realization

The circuit has been realized at the Delft Institute of Micro Electronics and Submicron technology (DIMES) in the DIMES01 process $(5GHz,\ 2\mu m)$, with vertical NPNs and lateral PNPs. Typical parameters for the NPNs are: $\beta_f = 100$, $f_T = 5GHz$, $V_{AR} = 4V$ and for the lateral PNPs: $\beta_f = 75$ and $f_T = 20MHz$. The capacitors are AlO_2 capacitors with a value of $0.36fF/\mu m^2$.

For the frequency compensation of the V_{be1} generator the following values proved to be enough: $R_{C1} = 10k\Omega$, $C_{C1} = 30pF$, $R_{C2} = 7k\Omega$ and $C_{C2} = 10pF$. For the pole-zero cancellation replacing a pole-splitting network in the V_{be2} generator, a resistor of $150k\Omega$ and a capacitor of $30pF$ showed to be sufficient. For the combiner the pole-zero cancellation network is implemented by a resistor of $10k\Omega$ and a capacitor of $40pF$. The resistive broadbanding is done by a resistor of $10k\Omega$. Finally, the PTAT source and constant current source are stabilized by, respectively, $60k\Omega$ with $10pF$ and $200k\Omega$ with $20pF$.

The resistors for the scaling factors are $32k\Omega$, $7.5k\Omega$, $25.5k\Omega$ and $14.4k\Omega$ for, respectively, R_{11}, R_{12}, R_{21} and R_{22}.

In Figure D.12 a photo of the chip is depicted. On this chip the resistors R_{11}, R_{12}, R_{21}, R_{22}, R_t and R_{PTAT} are made controllable for testing purposes.

D.6.3 Measurement results

The measured output voltage as a function of temperature is depicted in Figure D.13. Other measurement results are summarized in table D.1.

From calculations on the idealized bandgap reference, a minimum temperature dependency of 0.22 ppm/K can be found for the corresponding temperature range. This remaining dependency is a result of the non-compensated third and higher-order temperature dependencies of the base-emitter voltage. However, to reach this, the influence of the remaining of the implementation must be negligibly small.

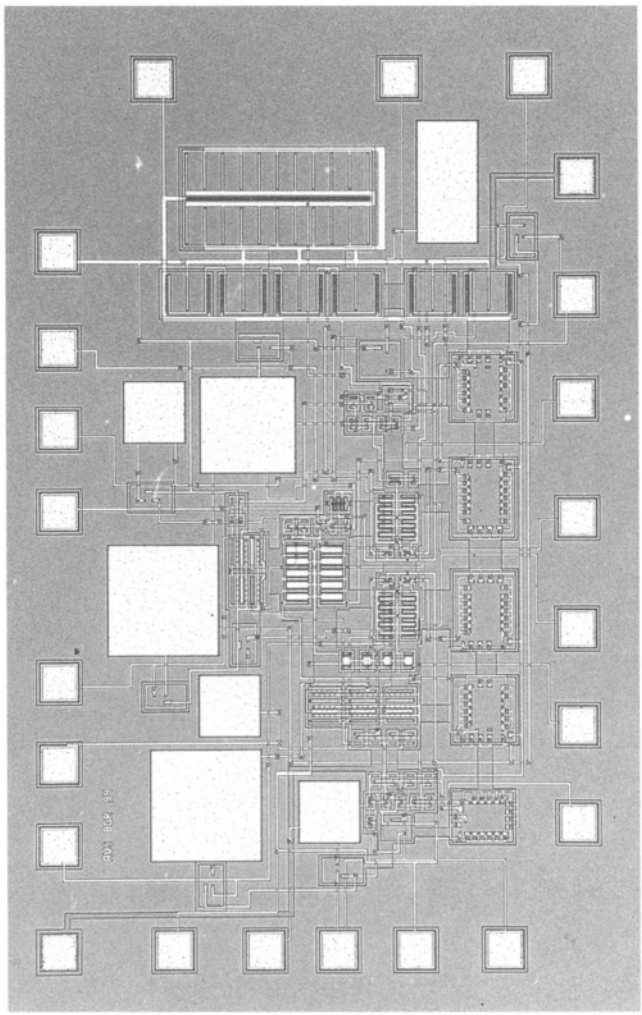

Figure D.12: A chip photo of the bandgap reference

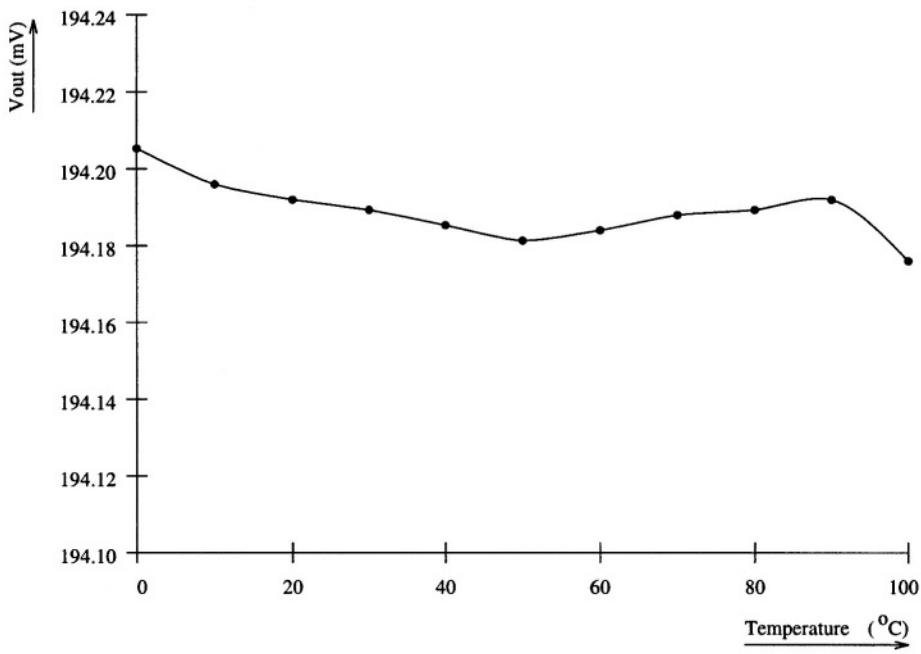

Figure D.13: The measured output voltage as a function of the temperature

Output voltage (mV)	194.2
Mean temperature dependency [0 °C-100 °C] (ppm/°C)	1.5
Output impedance (V/A)	43
Supply current (μA) at T=27°C	101
Noise ($nV/\sqrt{(Hz)}$)	<80
Line regulation (dB)	-80
Total capacitance (pF)	150pF
Chip size (mm^2)	3

Table D.1: Performance of the realized circuit

The main causes for the deviation of the realized bandgap reference from this number are twofold. First, for the lower temperatures the voltage available for the tail-current source of the differential pair in the summator becomes too low. Consequently, saturation of this source occurs, and as the differential pair is not used completely symmetrically, errors are found in the reference voltage. Note that the 1.5 ppm/K over the complete temperature range of this bandgap reference is equivalent to a total deviation of only 30 μV, see figure D.13. This behavior can be improved by enlarging the voltage drop across the current source or by choosing summation in the current domain; the differential pair and the current source can then even be removed. The current sources supplying the currents for the two reference transistors do not introduce errors as their voltage is at least about 250 mV at 0 °C and thus the errors are negligible, see section 6.7.2.3.

Second, at the higher end of the temperature range, the deviation is mainly caused by the influence of leakage currents. At about 125 °C a sharp drop in the reference voltage was found (on the order of several mV over a range of 10 °C), which influence is already noticeable at 100 °C, see figure D.13. For improving the bandgap reference in this region, the leakage currents and their temperature behavior have to be taken into account. In [4] a thorough treatment of the currents in the bipolar transistor can be found, including saturation currents (i.e. the leakage currents of PN-junctions).

This bandgap reference was specially designed to verify the feasibility of a temperature compensation by means of a linear combination of base-emitter voltages. At the time of the design, the the noise production of second-order compensated bandgap references was not yet studied in detail. For the bias current of the two reference transistors a large ratio was chosen as this was thought to be a correct choice. The total equivalent noise production of the idealized bandgap reference (sealer, adder and biasing still ideal) amounts to about 20 $\mathrm{nV}/\sqrt{\mathrm{Hz}}$. One reference transistor was biased at $\approx$15 μA whereas the other was biased at $\approx$0.2 μA. However, later on from noise minimization was found, see section 6.5.1.2, that for optimum noise performance the current ratio

of these two currents should be the same as the ratio of the two scaling factors (which equals about 0.8 and differs considerably from the used collector current ratio). When the optimum ratio for the collector currents is used, equation (6.65) applies and the minimum noise level for a current consumption of $15.2\ \mu\mathrm{A}$ is found to be $5.1\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$; about a factor 4 better. As the noise contribution of the biasing is a relative contribution, the expected noise of the complete optimized bandgap reference, for the same power consumption is about $20\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$ (from table F.1 a ratio of 4 is found between the noise of the idealized bandgap reference and the noise of the idealized bandgap reference). This is a factor 4 lower compared with the noise production of the realized reference!

D.7 Conclusion

In this appendix the design of a 1-V second-order compensated bandgap reference circuit has been presented. Only two base-emitter voltages are needed to obtain compensation up to the second order. One of the transistors is biased with a PTAT current and the other is biased with a constant current. The realized circuit has an output voltage of approximately 194mV and the mean temperature dependency is 1.5 ppm/$^\circ C$ in the range of 0$^\circ C$ to 100$^\circ C$. The circuit has been realized in a bipolar process with $f_t \geq 5GHz$. The total amount of capacitance is approximately 150pF and the current consumption is about $100\mu\mathrm{A}$.

Bibliography

[1] A. van Staveren, J. van Velzen, C.J.M. Verhoeven, and A.H.M. van Roermund. An integratable second-order compensated bandgap reference for 1V supply. *Analog Integrated Circuits and Signal Processing,* 8:69–81, 1995.

[2] E.H. Nordholt. *Design of High-Performance Negative-Feedback Amplifiers.* Elsevier Scientific Publishing Company, Amsterdam, 1983.

[3] H.C. Nauta and E.H. Nordholt. New class of high-performance ptat current sources. *Electronics Letters,* 21(9):384–386, April 1985.

[4] K. v.d. Lingen. *Bipolar Transistors for usage in Monolithic Bandgap References and Temperature Transducers.* PhD thesis, Delft University of Technology, Delft, 1996.

Appendix E

Optimum ratio of saturation currents

In this appendix the noise level of a first-order compensated bandgap reference is minimized, with the ratio of the saturation currents of the two reference transistors as the independent parameter. The noise of the first-order compensated bandgap reference can be described by:

$$S_v = 2q\frac{V_{REF}^2}{V_G'^2(0)} \cdot \left(\frac{T}{T_r}\right)^2 \times \frac{I_{C1}\left[V_G'(0) - V_T \ln\left(\frac{I_{C1}}{I_{S1}}\right)\right]^2 + I_{C2}\left[V_G'(0) - V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right)\right]^2}{I_{C1} \cdot I_{C2}\left[\ln\left(\frac{I_{C1}}{I_{C2}} \cdot \frac{I_{S2}}{I_{S1}}\right)\right]^2}. \quad \text{(E.1)}$$

In order to simplify the calculations, the relation between the two saturation currents is defined as:

$$I_{S1} = I_{S2}\exp(x). \quad \text{(E.2)}$$

Substitution of this expression in equation (E.1) and rewriting the resulting equation yields:

$$S_v = C \cdot \frac{I_{C1}\left[V_G'(0) - V_T \ln\left(\frac{I_{C1}}{I_{S2}}\right) + V_T \cdot x\right]^2 + I_{C2}\left[V_G'(0) - V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right)\right]^2}{I_{C1} \cdot I_{C2}\left[\ln\left(\frac{I_{C1}}{I_{C2}}\right) - x\right]^2}, \quad \text{(E.3)}$$

I_{C1}	I_{C2}	I_{S2}	$(I_{S1}/I_{S2})_{extreme}$
1 μA	10 μA	18 aA	7.66e-102
1 μA	10 μA	180 aA	7.66e-113
10 μA	1 μA	18 aA	7.73e-11
10 μA	1 μA	180 aA	6.14e-12

Table E.1: The optimum ratio of the two saturation currents for given collector currents and I_{S2}.

where C consists of the constant factors before the fraction of equation (E.1). The derivative of this expression with respect to x equals:

$$\frac{dS_v}{dx} = -2C\frac{\left[V'_G(0) - V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right)\right]}{I_{C1}I_{C2}\left[\ln\left(\frac{I_{C2}}{I_{C1}}\right) + x\right]^3} \times$$

$$\left[V'_G(0)I_{C2} + V'_G(0)I_{C1} - I_{C1}V_T \ln\left(\frac{I_{C1}}{I_{S2}}\right) + xI_{C1}V_T - I_{C2}V_T \ln\left(\frac{I_{C2}}{I_{S2}}\right)\right]. \quad \text{(E.4)}$$

The solution of $\frac{dS_v}{dx} = 0$ is given by:

$$x_{extreme} = -\frac{V'_G(0)}{V_T}\left(1 + \frac{I_{C2}}{I_{C1}}\right) + \ln\left(\frac{I_{C1}}{I_{S2}}\right) + \frac{I_{C2}}{I_{C1}} \cdot \ln\left(\frac{I_{C2}}{I_{S2}}\right). \quad \text{(E.5)}$$

Transforming this solution to the optimum ratio of saturation currents via:

$$y_{extreme} = \left(\frac{I_{S1}}{I_{S2}}\right)_{extreme} = \exp(x_{extreme}), \quad \text{(E.6)}$$

yields:

$$y_{extreme} = \frac{\left(\frac{I_{C1}}{I_{S2}}\right) \cdot \left(\frac{I_{C2}}{I_{S2}}\right)^{\left(\frac{I_{C2}}{I_{C1}}\right)}}{\exp\left[\frac{V'_G(0)}{V_T}\left(\frac{I_{C2}}{I_{C1}} + 1\right)\right]}. \quad \text{(E.7)}$$

In table E some examples are given of this extreme. For these examples it holds that: $V'_G(0) = 1.2$ V and $V_T = 25$ mV. Clearly, the extreme is located at a very small ratio of the saturation currents. Now it has to be checked whether this extreme is a minimum or a maximum. For this the sign diagram is derived. The extreme given in equation (E.7) is the only zero of equation (E.4). Further, equation (E.4) has a triple pole at:

$$x_{triple-pole} = \ln\left(\frac{I_{C1}}{I_{C2}}\right), \quad \text{(E.8)}$$

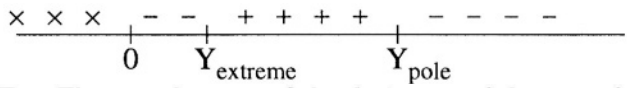

Figure E.1: The sign diagram of the derivative of the noise function.

corresponding to a triple pole of y at:

$$y_{triple-pole} = \frac{I_{C1}}{I_{C2}}. \tag{E.9}$$

This is again the condition for equal base-emitter voltages. In figure E.1 the sign diagram is depicted. For practical ratios of the collector currents, the zero is always closer to the origin than the triple pole. It is easily seen that for very large x and thus y, the derivative is negative, yielding the signs of the derivative as given in figure E.1. Thus the extreme found is a minimum and the triple pole corresponds to a maximum (which was already found before). Of course, the minimum is at a very impractical ratio. In practical cases, the ratio of the two saturation currents will be relatively close to the pole. As the pole corresponds to a maximum, the noise reduces for ratios going away from this pole. The remaining question to be answered is, do I have to choose a ratio that is larger or a ratio that is smaller than the ratio corresponding to the pole? That is easily found from figure 6.11. When $I_{C1}/I_{C2} > 1$, for the saturation currents the following must hold: $I_{S1}/I_{S2} < 1$. Thus the largest transistor gets the smallest current.

Appendix F

Design example: first-order compensated BGR

F.1 Introduction

In section 6.5.1.1 the noise behavior of an *idealized* bandgap reference was treated. In this appendix a design example of a *practical* low-noise bandgap reference [1] is described to determine which parts of the implementation contribute in what extent in addition to the noise, so that the fundamental limit is not reached.

First attention is paid to the basic structure of the bandgap reference and subsequently the constituting blocks are implemented one by one. Via simulations the noise performance is verified.

F.2 The basic structure of the design example

For this design example a simplified structure is used. This structure is obtained by shifting one scaling factor out of the bandgap reference (cf. section 6.8). In figure F.1 the block diagram of the example bandgap reference is given. The block diagram is found from figure 6.7 by shifting sealer a_2 through the summing node and subsequently deleting this sealer at the output. As the sealer a_2 is temperature independent the newly created reference voltage is also temperature independent ($V'_{REF} = V_{REF}/a_2$). Here it is assumed that sealer a_2 is negative and a_1 is positive. Thus the summing node has to subtract V_{BE2} from the scaled V_{BE1}. For this block diagram the noise optimization of the previous sections still holds. Scaler a_2 does not contribute any noise in the optimization. So, when the ratio of the reference voltage V_{REF} and the noise voltage at the

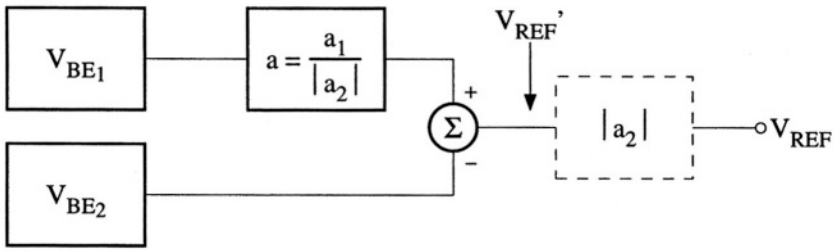

Figure F.1: Block diagram of a simplified first-order compensated bandgap reference

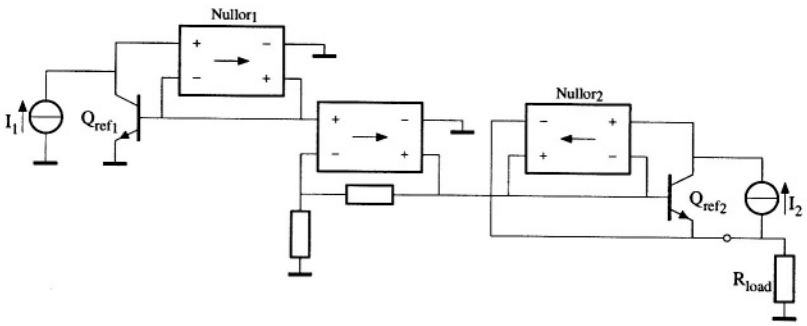

Figure F.2: Implementation of the bandgap reference at nullor level

output of a_2 (figure F.1) is optimal, the ratio at the input of a_2, V'_{REF}/v_{noise}, is also optimal.

With shifting out one of the scaling factors the number of degrees of freedom is reduced by one. However, as for the first-order compensated bandgap reference one degree of freedom less is used compared with the special situation for the second-order compensated bandgap reference, the reference voltage can still be freely chosen.

The remaining scaling factor is totally determined by the first-order compensation. The new scaling factor equals the ratio of the two previously used scaling factors

$$a = \frac{V'_G - V_{BE2}}{V'_G - V_{BE1}} \tag{F.1}$$

and when the two base-emitter voltages are given the reference voltage is fixed ($V_{REF} = aV_{BE1} - V_{BE2}$)

In figure F.2 the first step down to a total implementation is given. Transistor Q_{ref1} and Q_{ref2} are the transistors used for the generation of the two required base-emitter voltages. Nullor 1 forces the base-collector voltage of Q_{ref1} to zero in order to minimize the influence of the forward Early effect. Further, the input

current of the nullor is zero and thus the current from source I_1 flows completely through the collector lead of Q_{ref1}. Finally, the nullor buffers the base-emitter voltage such that load currents do not influence the base-emitter voltage. Nullor 2 performs the same as Nullor 1 does, but now for transistor Q_{ref2}.

For this example the scaling ratio of the two reference transistors is chosen to be 1:10. The resulting optimal current division follows from equation (6.59) as 1:0.28. When a current consumption of 1.25 μA is chosen the two collector currents needs to be approximately

$$I_{C_{ref1}} = 1\ \mu\text{A}, \tag{F.2}$$

$$I_{C_{ref2}} = 0.25\ \mu\text{A}. \tag{F.3}$$

The biasing point is chosen somewhat beside the optimum in order to get a convenient scaling ratio between the two bias currents. But as the noise-versus-current-division graphic is relatively flat, the influence up on the noise level is negligible (figure 6.11). For transistor 1 a transistor with a saturation current of I_S=16 aA is used [1]. Then the reference voltage equals to (V'_G=1.19 V):

$$V'_{REF} \approx 206\ \text{mV}. \tag{F.4}$$

The design is done for integration in the DIMES01 process [2], which is a bipolar process. In the next section the two base-emitter voltage generators will be implemented.

F.3 Implementation of the two V_{BE} generators

The implementation for the first base-emitter voltage generator is given in figure F.3. Although the differential pair seems to be the obvious choice for the input stage of the nullor implementation, it is not used here. As the voltage at the emitter nodes of the two transistors becomes relatively low, the implementation of the tail current source is hampered. However, when a CE-stage is chosen as an input stage this problem does not exist. But now the base-collector voltage of Q_{ref1} is in the order of 0.1 V. Fortunately, the error introduced via the forward Early effect is still negligible because of the relatively high forward Early voltage. A current mirror is used to obtain a negative loop gain. The noise introduced by the nullor implementation is negligible because of the gain of the reference transistor. Frequency compensation is obtained by the pole-splitting network R_{sp1} and C_{sp1}.

The implementation of the second base-emitter voltage generator is depicted in figure F.4. For this nullor implementation the drawback of the use of a dif-

[1] In this value the influence of a reverse Early voltage of V_{AR}=4 V is incorporated. $I_{S-with-V_{AR}} = I_{S-without-V_{AR}} \cdot (1 - V_{BE}/V_{AR})$. For small variations in V_{BE} the last factor can be assumed to be constant.

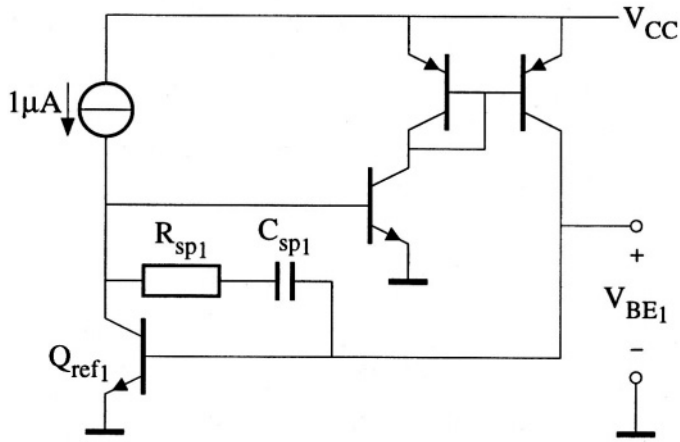

Figure F.3: Implementation of the first base-emitter voltage generator

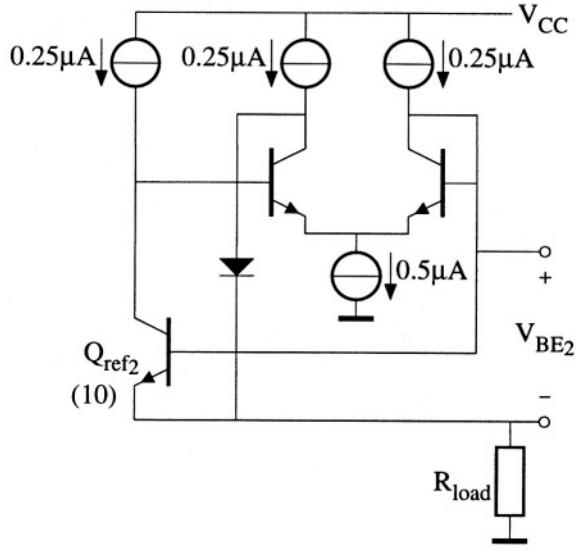

Figure F.4: Implementation of the second base-emitter voltage generator

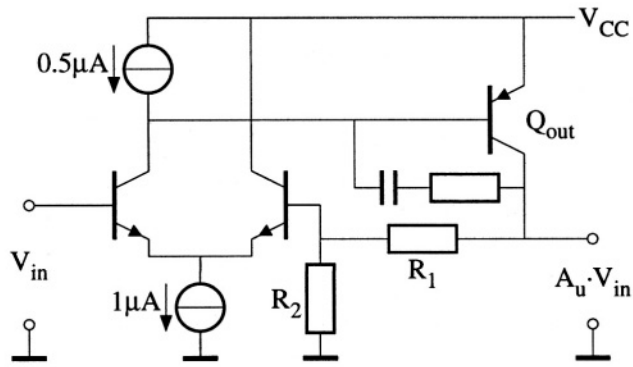

Figure F.5: The negative-feedback voltage amplifier

ferential pair as input stage is not apparent because the common-mode voltage of this nullor is 200mV higher than it is for nullor 1. As the differential pair has a non-inverting output, no additional stage is needed to obtain a negative loop gain. Without additional frequency compensation components the circuit already showed an acceptable frequency behavior. As for the V_{BE1} generator, the noise is predominantly determined by the collector shot noise of the reference transistor.

At this point of the design the total noise power density of the (idealized) bandgap reference equals:

$$S_{output} = 2kT(a^2 r_{e1} + r_{e2}) = 34^2 \text{ [nV]}^2/\text{Hz} \tag{F.5}$$

with $r_e = kT/qI_C$. In the next section an implementation shall be made for the scaler.

F.4 The implementation of the scaler

In figure F.2 the scaler was implemented as a negative-feedback voltage amplifier. This voltage amplifier is depicted again in figure F.5.

The voltage gain A_u is set by the ratio of the two resistors as

$$A_u = 1 + \frac{R_1}{R_2}. \tag{F.6}$$

The input offset voltage of the nullor implementation is directly added to the reference voltage. So, this offset voltage must be as small as possible. Consequently, for the input stage a differential pair is chosen.

The output of the scaler has to supply for the current through the feedback network and the input current of the V_{BE2} generator. The current through

the feedback network is related to V_{BE1} and thus temperature dependent. The input current of the V_{BE2} generator equals its load current, as the generator acts as a floating voltage source, plus an input offset current. This input offset current is also temperature dependent. Thus the total load current of the scaler is temperature dependent. Transistor Q_{out} is used to supply this load current and to reduce the influence of this current on the input offset voltage.

The equivalent input noise power density of the scaler equals

$$S_{scaler} = 4kT(R_1//R_2 + r_e) + 2qI_B \cdot (R_1//R_2)^2. \tag{F.7}$$

The first term is due to the thermal noise of the feedback network and the equivalent voltage noise of the input stage. The second term is due to the equivalent noise current of the input stage. A minimum exists for the noise contribution of the input stage. For increasing collector current the equivalent voltage noise of the input stage decreases, but the influence of the equivalent current noise increases, and vice versa for a decreasing collector current. Thus a minimum is obtained when for a small change in collector current the change in equivalent noise voltage of the input stage is compensated for by the complementary change of the equivalent input noise current. As the influence of the input noise current is dependent on the feedback resistors, this minimum is too.

For the noise due to the feedback resistors it holds that the lower the resistor values are, the less the noise contribution is. But as the voltages across those resistors are determined by the base-emitter voltages, the current consumption is directly related to their noise performance. For a lower noise contribution a higher current consumption is required.

For the current consumption of the resistors as a function of the base-emitter voltages it holds that

$$I_{fb} = \frac{V_{BE1}}{R_1} \tag{F.8}$$

in which I_{fb} is the current through the feedback network. The influence of the equivalent input noise voltage of the scaler on the noise at the output of the reference is found by multiplying by the scaling factor a. As the scaling factor is given by

$$a = \frac{V_G' - V_{BE2}}{V_G' - V_{BE1}} \tag{F.9}$$

and the practical base-emitter voltages are limited to a relatively small range, the factor is more or less independent of the base-emitter voltages.

Concluding, the noise contribution due to the feedback resistors can only be reduced, reasonably, by increasing the current consumption. A compromise has to made between the current consumption and noise contribution. For the feedback resistors the following values are chosen

$$R_1 \quad = \quad 500\ \text{k}\Omega, \tag{F.10}$$

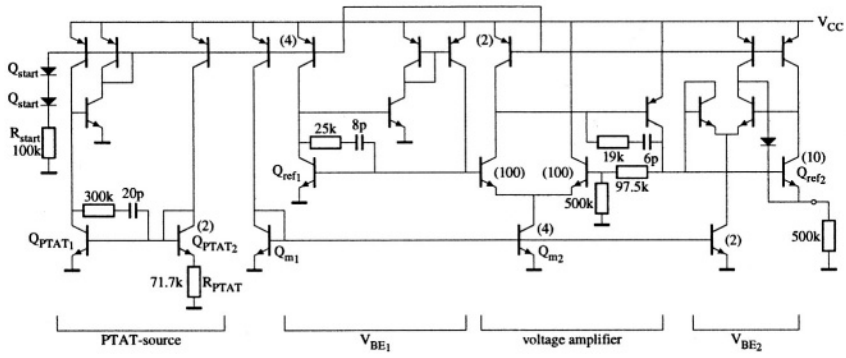

Figure F.6: The total circuit of the bandgap reference

$$R_2 = 500\ \text{k}\Omega * (1 - \text{a}) \approx 86.2\ \text{k}\Omega, \quad \text{(F.11)}$$

$$I_{fb} \approx 1.3\ \mu\text{A}. \quad \text{(F.12)}$$

Resistors of this value are readily available in current technologies. Because the resistors set an amplification factor, only the matching is important. For these resistors the current consumption is approximately equal to the current consumption of the two reference transistors. The noise contribution is of the same order.

Now that the feedback resistors are known, a noise minimization for the input stage can be performed. Doing so, an optimal collector current of 3.5 μA for each transistor of the input stage is found. In that case the equivalent noise resistor of the input stage amounts to 15 kΩ, which is negligible. But the current consumption is relatively high and thus a lower collector current is chosen: 0.5 μA for each input transistor. Now the equivalent noise resistor equals 100 kΩ and the noise contribution is of the order of the noise due to the V_{BE2} generator.

The high-frequency behavior of the sealer is compensated by a pole-splitting network.

F.5 The complete circuit

Now al the parts have been implemented, the bias circuitry is designed. The total schematic of the bandgap reference including bias circuitry is depicted in figure F.6.

The bias currents are referred to a PTAT current generated by the difference of two base-emitter voltages Q_{PTAT1} and Q_{PTAT2}, together with R_{PTAT}. The

noise contribution of this part to the noise of the bias currents is given by

$$S_{iptat} = \frac{4kT}{R_{PTAT}} \left(\frac{r_{e_{ptat1,2}}}{R_{PTAT}} + 1 \right). \tag{F.13}$$

To realize a negligible contribution to the noise, the currents in the PTAT source need to be relatively large ($r_{e_{ptat1,2}} \ll r_{e,Qref1,2}$) and a large difference in the base-emitter voltage is needed in order to be able to use a high value for R_{PTAT}. In contrast, the noise due to the current mirrors on the top of the PTAT source and the transistors implementing the current bias sources can not be made negligible because of the power supply voltage of only 1 V. For this design example the 1 V power supply voltage is a constraint. Thus lowering the influence of the noise of the PTAT current source must be done by increasing its current consumption.

As a compromise between current consumption and noise contribution, the following values are used for the PTAT source

$$I_{PTAT} = 0.25\ \mu\text{A}, \tag{F.14}$$

$$\Delta V_{BE,PTAT} = 18\ \text{mV @ 300 K (scaling 1 : 2)}, \tag{F.15}$$

$$R_{PTAT} = 71.7\ \text{k}\Omega. \tag{F.16}$$

Startup is secured by the two diodes $\mathrm{Q_{start}}$ and a resistor R_{start}.

A simulation result of the temperature behavior of the total bandgap reference is given in figure F.7. The mean temperature dependency equals 20 ppm/K for 0 °C to 100 °C and the current drain is $\approx 5\ \mu\text{A}$ from a 1 V power supply. The output voltage differs slightly from 206 mV as was calculated before. This is caused by the trimming that was required because of a small temperature-dependent input offset voltage of the voltage amplifier. This trimming resulted in a small change of the nominal value. The cause of the small input offset voltage is twofold.

Firstly, the influence of the load current of the scaler is not totally negligible. This influence can be reduced by adding a third amplifying stage. But problems can be expected with the frequency compensation.

Secondly, the mismatch between the tail-current source and the current from the PNP current source of the differential pair is such that it results in a non-negligible offset voltage at the input. This mismatch is predominantly caused by the error due to the base currents in the mirror factor of the NPN mirror $\mathrm{Q_{m1}}$ and $\mathrm{Q_{m2}}$. This error can be reduced to a negligible level by connecting $\mathrm{Q_{m1}}$ via amplifier stages as a diode (compare to the reduction of the influence of base-currents in the V_{BE1} generator). But as this bandgap reference is only a demonstrator for low-noise design, these measures are not taken.

In figure F.8 a simulation result of the total noise contribution of the bandgap reference is depicted.

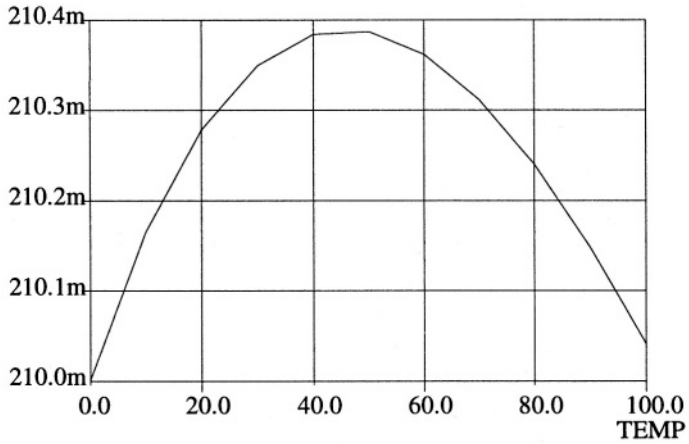

Figure F.7: Simulation of the temperature behavior of the total bandgap reference

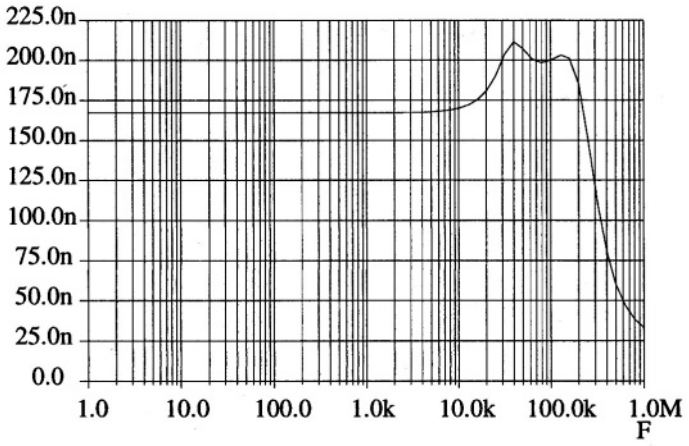

Figure F.8: Simulation of the noise at the output of the total bandgap reference

The total noise amounts to approximately $166\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$. The contributions of the different parts of the bandgap reference are given in table F.1

Part	Contribution
V_{BE1} generator	$15\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$
V_{BE2} generator	$30\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$
Voltage amplifier	$54\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$
Biasing circuit	$153\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$
Total	$166\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$

Table F.1: The noise contributions of the different parts of the bandgap reference

Note that the noise contribution of the biasing circuit is relatively large. This is inherent in low-voltage design (in this case 1 V). All the transistors used for the biasing contribute at least $2qI_C$ noise to the bias currents and the PTAT current source makes an additional contribution to the noise. Noise contribution can be minimized when emitter resistors are used. With the emitter resistors, the noise contribution can be reduced to $2qI_B$. However, to reach this, relatively large resistors and thus a relatively high voltage for a given current, is needed. It can be calculated fairly simple that in order to obtain a noise contribution of $4qI_B$, which is still negligible, the voltage across the emitter resistors needs to be approximately 5 V. A closer look shows that a considerable reduction of the noise contribution is already obtained for about 100 mV across the resistors. Then the noise power of the biasing can be reduced to about 5 % of the original level, see equation (6.78).

Thus one can say that the performance versus power consumption of bandgap references is not degraded that much by the 1 V power supply constraint. For 1 V design, the influence of the bias sources on the noise behavior can be accounted for by just a factor. Of course this factor depends on how the bias circuit is designed.

The bias circuit not only contributes to the noise by its own shot noise. Noise from the power supply penetrates through the *practical* bias sources to the output of the bandgap reference and contributes to the noise as well. The sensitivity of the bandgap reference to power supply noise is determined by the implementation for the bias sources (how much noise is injected) and the transfer of the injected noise to the output (what is seen of the injected noise). At relatively high frequencies in particular it is hard to make a good implementation of the bias sources such that the injected noise is kept low. When the injected noise is predominant at high frequencies, additional measures in the bandgap reference circuit have to be taken such that noise injected at different nodes cancel at the output of the bandgap reference or are attenuated in the bandgap reference. These measures can be done independently of the noise optimization.

Besides the noise contribution due to the bias circuit, the noise of the scaler prevents the bandgap reference from reaching the fundamental limit. As was pointed out, more current through the feedback resistors and input stage reduces the noise contribution of the scaler. But the noise of the idealized bandgap reference (the two base-emitter voltage generators) also decreases when larger currents are used. Thus in the division of the total current between the two base-emitter voltage generators and the scaler there is an optimum at which the total noise level is minimal. The noise optimization described by the strategy in this appendix will not be far from this global optimum. This is because the noise performance of the scaler is only slightly influenced by the values of the base-emitter voltages of the reference transistors [via the scaling factor equation (F.1)]. Thus when noise of the two base-emitter voltages and the scaler are minimized separately and their levels are comparable, the total noise level will be close to the global optimum.

F.6 Conclusions

In this appendix a design example was shown of a special structure of a first-order compensated bandgap reference. For this design noise minimization was the key issue, i.e. how close can the noise level be to the noise level of the idealized bandgap reference.

For this example it was found that the noise contribution of the voltage amplifier, implementing the single scaling factor, is easily made on the same order of magnitude as the noise contribution of the idealized bandgap reference. On top of that it was shown that the noise contribution due to the biasing circuitry is easily made small compared with the noise of the rest of the reference.

The designed bipolar bandgap reference has an output voltage of about 200 mV and the mean temperature dependency is ≈ 20 ppm/K for 0 °C to 100 °C (which is a direct consequence of the first-order temperature compensation). The output noise density equals $166\ \mathrm{nV}/\sqrt{\mathrm{Hz}}$. The total current consumption is $5\ \mu\mathrm{A}$.

Bibliography

[1] A. van Staveren, C.J.M. Verhoeven, and A.H.M. van Roermund. The design of low-noise bandgap references. *IEEE Transactions on Circuits and Systems I,* 43(4):290–300, April 1996.

[2] L.K. Nanver, E.J.G. Goudena, and H.W. van Zeijl. DIMES-01, a baseline BIFET process for smart sensor experimentation. *Sensors and Actuators Physical,* 36(2):139–147, 1993.

Index