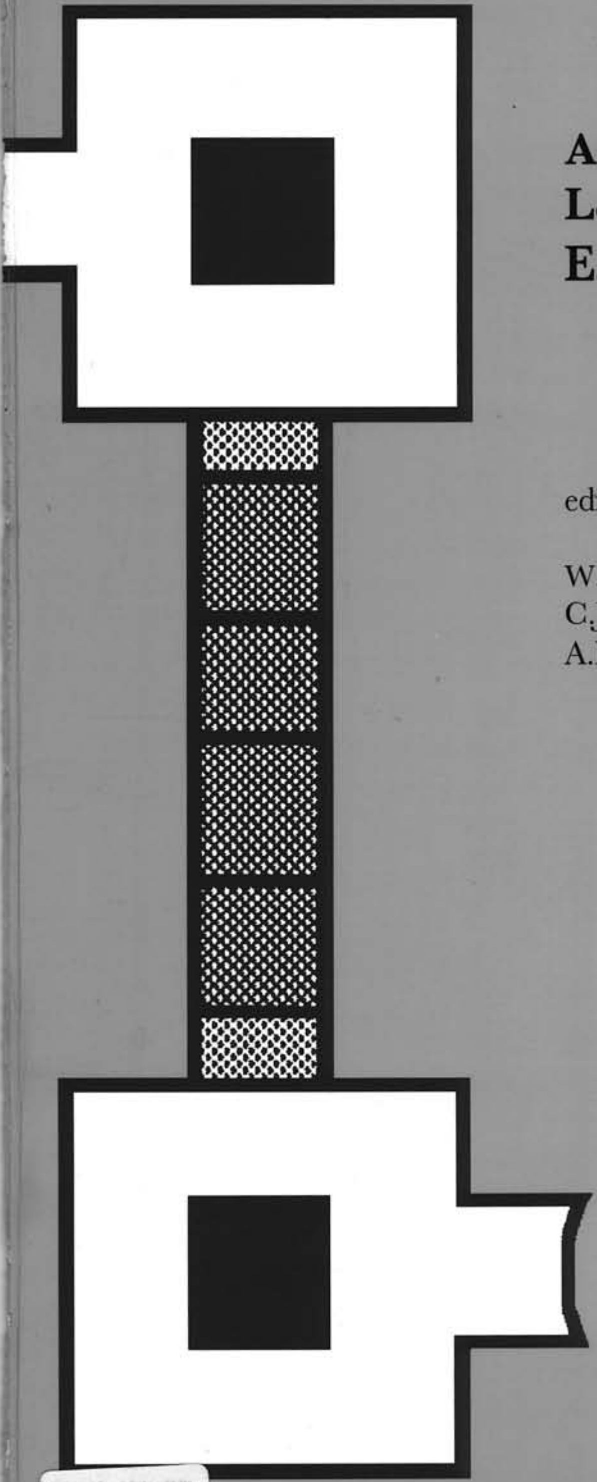


Analog IC Techniques for Low-Voltage Low-Power Electronics

edited by

W.A. Serdijn,
C.J.M. Verhoeven &
A.H.M. van Roermund

Analog IC Techniques for Low-Voltage Low-Power Electronics

edited by

W.A. Serdijn,
C.J.M. Verhoeven &
A.H.M. van Roermund

Delft University Press/1995

Published and distributed by:

Delft University Press
Stevinweg 1
2628 CN Delft
The Netherlands
Telephone +31 15 278 3254
Fax +31 15 278 1661

By order of:
Delft University of Technology
Faculty of Electrical Engineering
Electronics Research Laboratory
Mekelweg 4
2628 CD Delft
Telephone +31 15 278 6180
Fax +31 15 278 5922

CIP-DATA KONINKLIJKE BIBLIOTHEEK, THE HAGUE

Serdijn, W.A.

Analog IC Techniques for Low-Voltage Low-Power Electronics / W.A. Serdijn, C.J.M. Verhoeven and A.H.M. van Roermund. - Delft : Delft University Press. - Ill.
With ref.
ISBN 90-407-1202-6
NUGI 832
Subject headings: Analog Electronics / Low-Voltage / Low-Power

Printed in the Netherlands

to Jan Davidse

Contents

Chapter 1

Introduction

Arthur van Roermund

1.1 Electronics in a historic perspective

Electronics covers a rather broad area in the world of technology. Starting from discrete electronics with devices like radio tubes, it has evolved in the past decades to what is called micro-electronics: electronics in integrated circuits, with very small dimensions and with opportunities for very low-cost mass production. This process started with the invention of the transistor in 1947, and an essential breakthrough came with the development of a microtechnology that made it possible to integrate a number of transistors together on one substrate, the chip. As the process of miniaturization evolved, transistors have ever since shrunk, chip size increased, and the number of devices per chip has increased every year.

Handling the complexity of such large chips and the complexity of the design was, and is still, very difficult. The rise of digital control, digital data processing and, finally, digital signal processing has provided the means to abstract and to formalize the design process, to decouple developments in IC processing technology, developments on circuit level, and to developments on system level in order to cope with restricted accuracies in the processing by the electronic devices on the chip, and to cut the large system design problem into smaller parts that can be solved separately. All this together led to an unprecedented advance of the miniaturization. Terms as medium-scale integration, large-scale integration are already old-fashioned; even the concept of very-large-scale integration (VLSI) seems to have become inappropriate; today we would rather talk about super VLSI or ultra-large-scale integration (ULSI).

What was the place of analog electronics in all this? Of course, it all started with analog, but, as said, large parts of analog signal processing were taken over by digital signal processing. Analog signal processing, however, maintained its position in special circuits with high requirements, for example on power dissipation, and at the interface of "digital" chips with the outside world. However, analog electronics is more than analog signal processing: it is the implementation of a function in IC hardware with precise voltage and current waveforms. In fact, there is no other electronics than analog electronics: also digital circuits are in essence analog. The aforementioned evolution of the digital circuits, however, was not driven by improvements at the basic electronic level, but mainly at the IC processing level, and at the higher circuit and system levels. The growth in analog circuit design was just driven by the need for interface circuitry for the increased number of digital designs. The quality of the combined systems, however, was, and is still often, dominated by the accuracy of the analog interface circuitry.

1.2 Current trends in electronics

Today, (analog) electronics can rejoice in disproportionately increasing attention. Optimization at the device and circuit levels pays off again; several arguments can be given for this.

First, there is a enormous drive toward low-power and low-voltage ICs, both for analog and for digital processing. In this book, the low-power low-voltage aspect will get attention in all chapters. The reasons for the drive for low-power and low-voltage ICs are manifold, and will be discussed in more detail further on.

Second, the dimensions of the devices are becoming so small, that several "second-order" effects cannot longer be neglected. We can think of short-channel effects, but, in future, things like quantum effects will play a role. The modeling is becoming increasingly complex and the design is becoming really an analog design, even if the information is digital.

Third, the ICs are boosted to operate at higher frequencies. Digital circuits are already operating at 100MHz and even higher; their electrical waveforms have become analog. Analog telecom circuits on IC operate at 1 or 2 GHz, and, in special processes, up to tens of GHz. Technology is already offering heterogeneous bipolar silicon-germanium devices in "normal" silicon processes with transit frequencies in excess of 100 GHz.

Fourth, attention is paid again to analog signal processing at the system level, to achieve larger amounts of processing power on chip; the neural networks are especially good examples of this. The self-learning aspect of the signal processing carries the potential to overcome the problem of limited accuracy and to allocate there processing power where it is needed.

1.3 Low power and low voltage

As said, there is an enormous drive toward low-power and low-voltage ICs, both for analog and for digital processing, and the reasons for this are manifold. In fact, two objectives are discussed here as if they were one: "low power" and "low voltage". The reason why these are addressed in combination further on is that, in many situations, both are required. However, it must be emphasized here that this does not hold for all situations; sometimes only one of them is required, as will become clear from the following discussions.

1.3.1 Reasons for low power/low voltage

The best-known argument for low power/low voltage is that more and more electronic equipment must be portable; now that the equipment has become smaller by means of miniaturization, size is no longer an impediment to portability, and this means that the equipment must be battery operated. Also for reasons of size, only one single battery is strived for, which means operation at about 1 volt with minimum power dissipation, in order to enhance battery life.

"Small" can also be an argument in itself, if the size is limited by the application, as in hearing aids that must be worn inside the auditory canal, or if size is a selling argument, or if EMC requirements demand for small sizes as, for instance, in medical electrode acquisition and transmission circuitry. EMC, however, also demands low emission levels, which means low currents (especially in digital circuits), and thus low power.

Low cost can also be a driving force in the considerations of small-sized equipment with a low battery cost.

The argument of maintenance cost is less well known, but it is especially important in situations where the equipment is not easily accessible, so that replacement is difficult and/or expensive. Here, lifetime is a major point.

Apart from battery lifetime, there is another ground for low power dissipation, namely operation temperature, which influences the operation or even makes operation impossible. In very large processing chips, dissipation, thus temperature rise, is becoming a dominant obstacle which limits further integration density. Operation temperature can also influence the lifetime and reliability of the electronics in a negative way. Leakage currents increase rather fast with rise in temperature.

Scaling down chip dimensions, whereas breakdown voltages are kept the same, means lower supply voltage. This is what is going on in digital circuits: starting from about 15 volts in former days for CMOS logic and 5 volts for TTL logic, the norm is at this moment 3 volts; and it is expected that this specification will also go down. Also here, there is a direct relation to reliability and life time.

Safety can also be a reason for battery operation, and thus for low power/low voltage.

Finally, environmental arguments (green electronics) push the low-power development, especially for mass products like consumer electronics, as batteries are known as important contributors to pollution.

1.3.2 A definition of low power/low voltage

From the foregoing discussion, we can deduce that low voltage means about 1 volt for analog signal-processing electronics (battery voltage) and 3 volts for digital circuitry (breakdown voltage). Thus, the soft term "low voltage" can rather precisely be translated to a more rigid specification. Further decrease of supply voltage can be foreseen in future, possibly up to a few kT/q, but, at the moment, there is no real demand for it. The low power argument is, on the contrary, a rather soft requirement. Here, low power in fact means minimal power, and thus in the case of low voltage it means minimal currents. How low these currents will be depends on the other specifications; in most cases it might be better to speak of relatively low power, or power efficient, than of low power.

1.3.3 Low-power/low-voltage applications

During the discussion of the arguments behind low power/low voltage, some applications have already passed in revue. More examples of applications are discussed below, and in a more categorized way.

Portable consumer electronics form a large group; it comprises radios, televisions, telephones, portable computers, personal assistants, personal communication equipment, etc. Here, portability, lifetime, and pollution are the dominant incentives.

Bio-medical electronics form a second group, where size, accessibility, and safety are the primary motivations. To this group belong implantable and injectable electronic devices, as for instance pacemakers, active electrodes, cochlear implants, etc.

A third group is the cluster of electronics with difficult accessibility in oil rigs, satellites, smoke and gas detectors, etc. The service argument is dominant here.

Electronics in dangerous environments require low power dissipation for safety reasons; gauge meters in oil tanks, for instance, have to fulfill rather strict regulations in connection with explosion risks.

Temperature sensors, normally, are meant to measure outside temperature and not chip dissipation, so here also the incentive is the decrease of operation temperature, but now from a functional point of view.

Another application area is that of massive computing by very complex processor chips, where power dissipation (read "temperature") and supply voltage are both becoming a limiting factor for further progress in integration density.

1.3.4 A global approach to low power/low voltage

Low power and low voltage requirements are, in fact, in many cases derived from system-level requirements such as portability, size, low cost, lifetime, maintenance;

only in a few cases are they directly required by the IC, for example in the case of limited breakdown voltage or chip temperature. Thus, in general, the problem is at the system level. There, a distinction can be made between three types of power supplies: mains, batteries, and alternative supplies.

Much work is being done on improving batteries. Although this is not directly the field of an electronic designer, but more the field of physicists and chemists, it is very important to know what is going on in this field, as this will have direct impact on the derived requirements for the ICs. Characteristics that are the subject of research are larger capacity, better rechargeability, increased lifetime, less self-discharging, environment-friendliness, safety, and integration with packages or even chips.

Of course, also alternative supplies are looked at, and have already been looked at for a long time. Photo cells which convert light power into electric power are extensively used; but we can also think of power supplies based on heat differences, as is already seen in watches, or the irradiation of electromagnetic power, received by coils in, for instance, cochlear implants and proximity detectors. The concepts are already old, but they have only been used in a limited number of applications. However, now that electronics power dissipation has been decreased to very low levels, all these ideas are becoming feasible.

Another field of research is that of supply electronics and recharge electronics. For all kinds of supplies, it holds that the interface between the supply and the signal-processing electronics is done by supply electronics, which take care of the supply voltage control, short-circuit protection, up or down transformation, ac-dc converters, battery control for monitoring the battery status, etc. Research is done in all these fields, contributing also to the earlier-mentioned requirements at system level.

Last, but certainly not least, we have the field of low-power low-voltage electronics for signal processing, logic or data processing. Here also, we can look at the problem on various levels. For instance, we can look at filters that dissipate minimal power, given a certain requirement for the dynamic range and for the chip area. However, it might make more sense to relax the filter specifications by changing the system structure. In a radio receiver, this can be done, for example, by changing the detector circuit, or by using a different mixer. However, having a filter structure derived from the specifications originating from the system optimization, we can implement this filter in several ways. A well-known method is using state-space filters, and implementing the states with integrators, each integrator comprising an operational amplifier. These amplifiers, however, dissipate far more power than necessary; dedicated amplifier design will help significantly. Nevertheless, the optimization problem is seen as the optimization of two subproblems, whereas, at least theoretically, there is a better optimization if we optimize the total problem as a whole. As an example, we can take a notch filter. Here, the requirements for the amplifiers, translated from filter specifications, are only stringent at the notch frequency band; at other frequencies, the requirements are far more relaxed. This means that instead of making amplifiers with high bandwidths, we can better make amplifiers with exactly a ninety degree phase shift only at the frequency of the notch. This can be achieved by matching the impedances within the amplifier that are responsible for its time constants with those that are responsible for the integrator time constant. Thus, substantial power savings can be derived.

In all cases, however, we arrive at the level of basic circuits, like amplifiers, mixers, oscillators, bandgap reference circuits, and logic gates. Here, the problem can be countered very fundamentally, as discussed in subsequent chapters. The method proposed here to find the optimum solution with respect to low voltage and low power is based on a structured design approach that will be discussed briefly in the following sections.

1.3.5 Obstacles at basic-circuit level

Before entering into the design problem, it is good to have an overview of the basic problems that we encounter at the basic circuit level. First of all, a low-voltage supply limits the number of active devices that can be stacked; in the case of 1 volt, this is restricted in most cases to two transistors, for example, a common-emitter amplifying stage and an active current source. This is partly due to biasing requirements and partly because we still wish to have some voltage range left for signal swing, at least if we are working in the voltage domain. Working in the current domain makes life easier, as far as we observe the low-voltage requirement. However, we are not always free to choose current as our signal carrier. In, for example, integrated filters, the integration function is done by capacitors, which translate currents to voltages.

Low signal swing is directly related to the dynamic range, as noise is coupled to fundamental processes like shot noise and thermal noise. As a consequence, low power leads to less dynamic range. A similar argument can be found for distortion; decreased signal range means that we are forced to use the active devices in large-signal operation, which increases distortion, and so again, the dynamic range is deteriorated.

Also other specifications are influenced by the low-power low-voltage requirement: reduction of power dissipation means reduction of currents. This leads to higher impedance levels, so to higher resistances, and lower transit frequencies. This in turn requires larger chip areas as well as high-frequency devices, and is therefore expensive. Moreover, the accuracy of the transfer function is influenced in a negative way by parasitic effects, like leakage currents.

Finally, higher impedance levels also translate to increased sensitivity to external noise sources, thus to an enlarged susceptibility, and, therefore to growing EMC problems.

1.4 The design dilemma

The foregoing section addressed aspects of what can be called the design dilemma. The specifications can be seen as boundaries in an N-dimensional space: one boundary for every specification item. Simultaneously, there is a fundamental limitation for every specification item. The design problem can now be considered as finding the optimal center point between these boundaries (design centering). As this is rather abstract, a simplified representation is here used to explain the dilemma; instead of the N-dimensional domain we consider a two-dimensional plane, as shown in Fig. 1.1. For a limited number of characteristics, namely low-voltage, low-power, costs, chip area, speed and accuracy, both the fundamental and the specification boundaries have been visualized as straight lines; the optimum of the design must be found somewhere between the two hexagons. However, tightening the low-voltage and low-power requirements means that the inner forbidden hexagon is enlarged, and that, consequently, the design range is reduced. Moreover, the boundaries are coupled, which means that a shift in one specification directly influences the other specifications, resulting in a shifted hexagon. This interaction is increased with decreasing margins, which makes it increasingly difficult to split the overall design problem into subproblems. Consequently, the design complexity grows, more specialized designers are needed, and design time increases.

1.5 A low-power/low-voltage design approach

The design dilemma shows us that there are three types of boundaries: the fundamental limits, the specification limits, and the requirements with respect to the design process. A structured design process must therefore start with an evaluation of the boundaries in all these three areas. That is the "threats" side of the design problem: the boundaries.

Of course there is also another side of the picture; we not only have to look at what cannot be done, but also at options; the "opportunities" side of the problem. In other words: we must evaluate all kind of devices, combinations, etc. that can be used to solve design problems.

The next step is: structuring of boundaries and options, to bring some order to the chaos. This can best be done along the axis of the design trajectory, see Fig. 1.2. At the top level are the applications, describing what is really asked for; from this level the functional specifications are derived. Then, a system can be

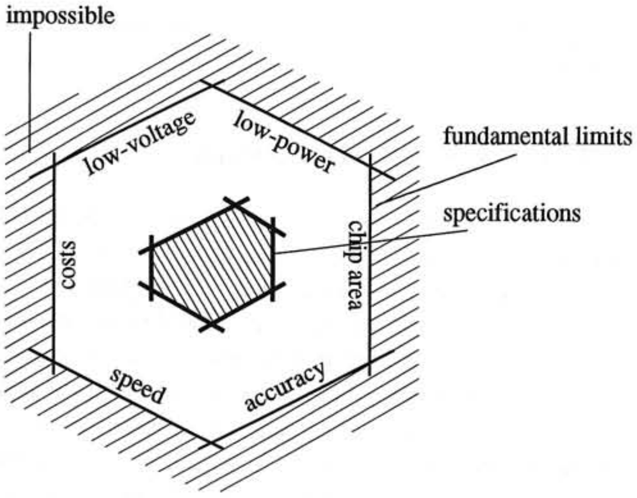


Figure 1.1: A two-dimensional representation of the design problem

designed at the high level, so let us say a block diagram based on the principal high-level functions that have to be realized and the relations between them. Each block in this diagram is further described on the basis of the basic functions in that block; for all blocks together, this leads to a collection of basic "primitive" functions that have to be realized: the functional level. Simulations at this level, and at the system level, are based on "behavioral modeling". For each primitive function, an algorithm, based on primitive operations, must be found that shows at an algorithmic level how the function will be implemented. This is still a completely abstract level of description, which, for instance, can be translated to a computer program, to simulate the algorithm. For the required hardware implementation, a different translation ("mapping") is required. For a computer program, for instance, normally a completely serial approach, described with finite loops, is followed. For the hardware, a combination of serial and parallel paths is used, which is normally shown in a signal-flow graph. The order in which all basic operations (for instance integrations, multiplications and additions in a filter) are executed is fixed now.

To here, it has all been abstract. The translation to hardware is made by designing circuits that can be assigned to the basic functions. Here, the electronic-circuit design takes place, starting with a mathematical function, and ending up with circuits built up with devices and interconnected according to the signal-flow graph, and a layout which adds the geometrical dimensions and positions for the combined circuit. After finishing this part of the total design, the IC is processed in a technology, with certain properties, like clearances, typical transistors, etc. Some of the properties of a technology are directly related to the materials used.

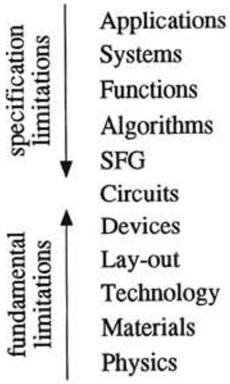


Figure 1.2: The various levels in the design trajectory, from applications to physics

And, finally, all processes taking place in these materials have to satisfy the laws of fundamental physics.

The reason why all these levels are mentioned here is that both limitations and possibilities are associated with each of these levels. Thermal noise is a fundamental physical process with fundamental limitations. Noise power in an RC circuit is related to the capacitance value. Component values are related to the circuit requirements, which are derived from the signal-flow graph, etc. Breakdown is a physical process, following fundamental physical rules. The field strength at which breakdown takes place depends on the properties of the materials used; breakdown voltages are dependent on these field strengths, on the dimensions in the layout, and on the voltages used at circuit level, etc. The velocity of electromagnetic waves is related to the material level, and it influences the time required for signal transport, and, therefore, also the circuit design, etc. It is clear that these kinds of limitations ripple upstream in the design trajectory. For the limitations forced by the specifications the reverse is true: they ripple downwards in the trajectory; they are imposed by the specific design requirements.

Structured design means, on the one hand, bringing structure to all the limitations and boundaries by ordering them to the various abstraction levels in the design trajectory, and, on the other hand, in a design methodology that brings a hierarchy to the design process, so that all limitations and possibilities are taken into account at the right moment, with design models that incorporate all relevant parameters for that level, but nothing more. Things that can still be done later on in the design trajectory must be done later on, and can, for the moment, be "forgotten".

Designing for low voltage and low power means following a structured design methodology that particularly takes into account the voltage and power limitations and options.

1.6 Ultra-low power

Low power and low voltage means voltages in the order of 1 volt and currents which are dependent on the application, but which are relatively low, and which go down to currents in the order of micro-amperes or nano-amperes; and corresponding power supply dissipation in the order of micro-watts or nano-watts. Here, some critical limits have been reached, for instance, the maximum resistance value that can be made on chip, or the minimum current through a transistor for which the current models are valid. Such limits are not fundamental, but they are difficult to pass, as new opportunities at circuit level, device level, etc. have to be found to overcome them. Technology can be forced to change (design-driven technology); new devices can be necessary, new and better modeling, etc. Examples of new devices are the single-electron transistor, the principles of which have been shown lately by physicists, and the silicon-germanium heterogeneous transistor. If the aforementioned currents and supply dissipations can be further decreased substantially, we speak about ultra-low power. This could pave the way for the alternative power supplies mentioned earlier.

1.7 Conclusions

Electronic design is part of a larger system-design problem and must be seen in that context. Low-power and low-voltage specifications are today gaining relatively more attention than the other design specifications, as certain boundaries are approached, and the voltage and power dissipation aspects are becoming crucial parameters for further improvements. The optimization problem can, and must, be encountered at all levels of the design trajectory and at all parts of the system. It involves the improvement of battery supply, investigation of alternative power supplies, better supply electronics, and better signal-processing electronics. Improvements in electronic design imply changes in circuit, device and technology approaches.

As we are coming increasingly closer to the boundaries, the design process is becoming increasingly complex and more difficult to split up into subproblems. Especially now, it is of utmost importance to find optimal design strategies in order to obtain optimal electronic circuits.

Chapter 2

Design principles

Wouter Serdijn

2.1 Introduction

An important criterion that must be fulfilled for all electronic systems is transfer quality. This quality is influenced by two different kinds of errors: stochastic ones and systematic ones. By stochastic errors we mean inaccuracies in the input-output relation caused by noise or interference. Though impossible to eliminate, their influence can be minimized by a proper design strategy.

Systematic errors arise from network imperfections, such as offset, non-linearity, inaccuracy of the device parameters, drift and temperature dependence. Probably the most effective method to reduce their influence, and thus to obtain an accurate transfer function, is by means of applying negative feedback, which allows us to exchange the large gain provided by the (highly non-linear) active devices for quality provided by (usually linear) passive devices.

Unfortunately, design strategies for obtaining a sufficient information capacity, or in general, transfer quality, (see, e.g., [1]) are normally not consistent with design strategies which take into account power dissipation, voltage range and current range. Therefore, it is the combination of transfer quality, low voltage and low power that must be considered during the whole design process.

In the following sections, attention is paid to five design aspects that all have an important influence on the overall system transfer quality: the system's input and output signals, the signal processing inside the system, the available technology, the parasitics and the power supply. It is shown that current becomes more favorable than voltage as the information-carrying quantity in a low-voltage low-power environment.

2.2 System requirements

The first step in the design process is to determine how the communication of the system with the external world, carried out by the source at the input and the load at the output, must be performed. Source and load are generally formed by other electronic systems or transducers. Depending on these, current or voltage (or a combination of these two), or linearly related quantities, such as charge or flux, must be chosen on the basis of the best reproducing relation to their physical input or output quantity [1].

The importance of choosing the correct source and load quantities can be illustrated by an example: a piezo-electric pressure transducer. A piezo-electric transducer, such as a piezo-electric microphone, converts pressure into charge. Since charge is linearly related to current ($i = dq/dt$), the output current of the sensor must be chosen as the electrical input quantity of the amplifier. The result is a charge amplifier. Yet, for many decades, voltage was chosen as the information-carrying quantity — the amplifier being a voltage amplifier — which caused the piezo-electric microphone to be considered inferior to its magneto-dynamic counterpart.

2.3 Signal processing

Assuming that the input and output quantities of the system have been determined by the foregoing system requirements, the next step in the design process is to decide which electrical quantities are best suited for a particular signal-processing function inside the system. When, e.g., signals coming from several subcircuits with a common terminal have to be added, current is a better choice for the information-carrying quantity than voltage. Currents can be added by simply connecting the output terminals of the subcircuits in parallel. When, however, a signal has to be distributed to several subcircuits, voltage is a better choice for the information-carrying quantity than current. Voltages can be distributed by simply connecting the input terminals of the subcircuits in parallel. For this reason, most of today's measurement instruments communicate by means of voltages, not currents.

Another example of choosing the correct electrical quantities is the use of a simple bipolar transistor when an exponential function over a wide range is required. Since the collector current is proportional to the exponent of the base-emitter voltage over a large range of collector currents, one device can do the job, if we are willing to choose voltage at the input and current at the output as the information-carrying quantity.

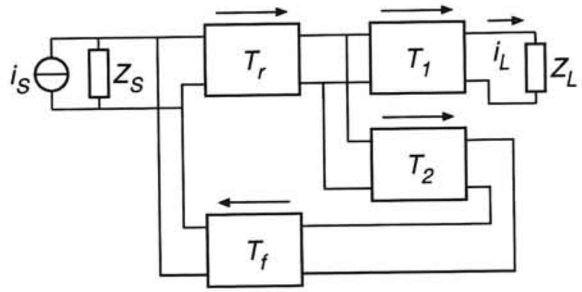


Figure 2.4: A current amplifier with negative feedback and indirect current sensing

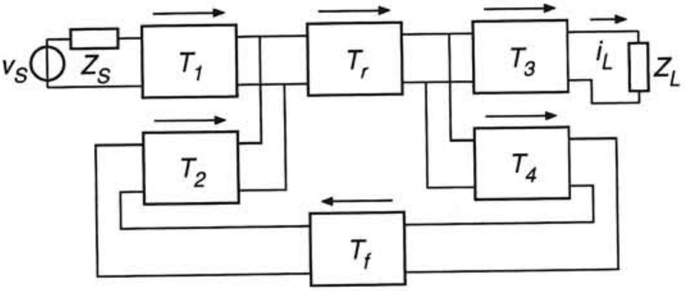


Figure 2.5: A transconductance amplifier with negative feedback and indirect current sensing and indirect voltage comparison

and $T_4 = T_3$, the output signal (v_L or i_L) is related to the input signal (v_S or i_S) as the inverse transfer function of the feedback network T_f.

As an example of the advantage of using indirect negative feedback in low-voltage circuits, let us consider the configuration shown in Figure 2.4, of which a possible embodiment is given in Figure 2.6. Two cascaded transistors inside T_r perform the nullor function and the feedback network is implemented by the resistive divider (R_1 and R_2). The indirect outputs are provided by T_1 and T_2. Clearly, now the maximum output voltage swing, and therefore the maximum output current swing, is limited only by the supply voltage (not shown) and the voltage across the output port of T_1 (the collector-emitter voltage). Similar arguments hold for the configurations of Figure 2.3 and 2.5.

2.5 The available technology

As third step in the design process, we now investigate how applying indirect negative feedback relates to the choice of the electrical quantities inside the system. In electronic circuits, indirect voltage comparison results in a doubled power den-

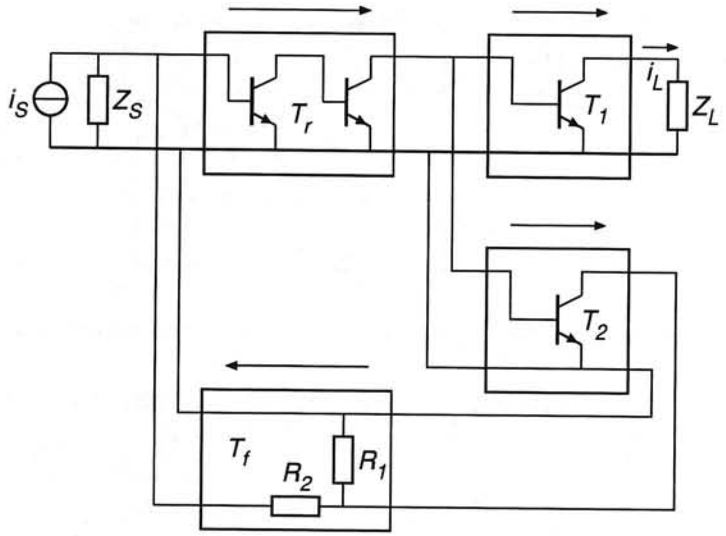


Figure 2.6: Possible implementation of the indirect-feedback current amplifier. Two cascaded transistors inside T_r perform the nullor function. The feedback network is implemented by the resistive divider (R_1 and R_2). The indirect outputs are provided by T_1 and T_2.

sity spectrum of the equivalent noise voltage at the input, because the direct and indirect input are connected in series. Indirect current sensing results in a doubled power density spectrum of the noise current at the output, because the direct and indirect output are placed in parallel. In practice, often the noise is most critical at the input, so on that ground there may be a preference for current sensing and thus for current as the information-carrying quantity.

Another disadvantage of the use of voltage as the information-carrying quantity is that, when the circuits are 'voltage-driven', i.e., from a low-impedance source, the equivalent input noise voltage is predominantly the result of the input noise voltage of both input stages. For bipolar transistors and MOS transistors in weak inversion, this input noise voltage is inversely proportional to the bias (collector or drain) current, and thus, in order to obtain a low input noise voltage, these bias currents must be rather large. This, of course, is in sharp contrast with our low-power requirement.

When, however, the circuits are 'current-driven', thus with a high impedance, the equivalent input noise current is mainly determined by the input noise current of the input stage. Since the input noise current of bipolar transistors and MOS transistors in weak inversion, in first approximation, is proportional to the bias current, this calls for small bias currents, which is in line with the low-power

2.4 Indirect Feedback

As mentioned earlier, systematic errors can be reduced by means of negative feedback. Figure 2.1 shows the four basic ways of applying (single-loop) direct feedback by means of two two-ports. If all the transfer parameters of two-port H approach infinity, i.e., H is a nullor, the output signal (v_L or i_L) is related to the input signal (v_S or i_S) as the inverse transfer function of the feedback network T_f.

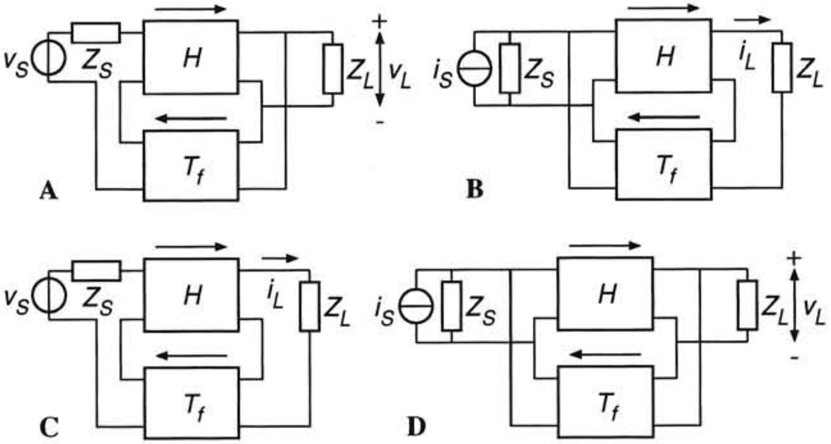


Figure 2.1: Four basic direct negative-feedback amplifiers: a voltage amplifier (A), a current amplifier (B), a transconductance amplifier (C) and a transimpedance amplifier (D)

In low-voltage circuits, however, due to the restricted voltage swing, it is often not possible, or at least not preferable, to connect two ports of these two-port networks in series, thus to sense the output current or to compare the input voltage of a circuit directly. This occurs in configurations A (at the input), B (at the output) and C (at both input and output). Hence, all direct-feedback configurations, except the transimpedance amplifier (configuration D), are less suited for low-voltage applications.

To clarify the disadvantage of connecting two ports in series, let us consider the configuration shown in Figure 2.1 B, of which a possible embodiment is given in Figure 2.2. Here, the two cascaded transistors Q_1 and Q_2 perform the nullor function while the feedback network is implemented by the resistive divider (R_1 and R_2). Clearly, the maximum output voltage swing, and therefore the maximum output current swing, is limited by the supply voltage (not shown), the voltage across the output port of the nullor (the collector-emitter voltage of Q_2) and the voltage across the input port of the feedback network (R_1). There are similar problems related to the low supply voltage for the configurations shown in Figure

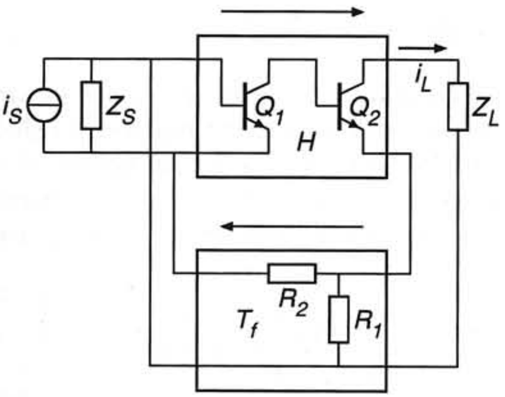


Figure 2.2: Possible embodiment of a direct-feedback current amplifier. Transistors Q_1 and Q_2 perform the nullor function. The feedback network is implemented by resistors R_1 and R_2

2.1 A and C.

To realize voltage, current and transconductance amplifiers, a useful alternative to direct negative feedback may be a technique called indirect negative feedback. In an indirect-negative-feedback circuit, the output and/or the input stage is copied, so that it has an equivalent input-output relation, and the feedback signal is taken from and/or fed back to that copy. Thus, it is possible to obtain a circuit response which is determined by the feedback network only, assuming that the copying does not introduce errors. A voltage amplifier, a current amplifier and a transconductance amplifier, all using the indirect negative-feedback principle, are depicted in Figures 2.3, 2.4 and 2.5. It can be seen that series-connected ports are now avoided in all configurations.

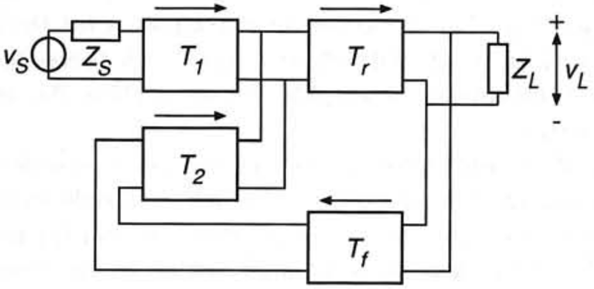


Figure 2.3: A voltage amplifier with negative feedback and indirect voltage comparison

Again, if all the transfer parameters of two-port T_r approach infinity, $T_2 = T_1$

requirement. This favors the choice of current as the information-carrying quantity.

A third disadvantage of indirect voltage comparison is that, in order to compensate each other, the non-linearities of the two input stages must be symmetrical or opposite, because the sum of their output currents must be nullified by the nullor. In practice, this requires either two balanced input stages or two complementary stages in a complementary IC process. The use of two balanced input stages, since their input noise voltages are placed in series, again doubles the power density spectrum of the equivalent input noise voltage. A complementary IC process is often not available and, moreover, exact complementarity can never be accomplished.

Indirect feedback at the output, however, calls for two identical output stages, to compensate for the non-linearities. These can easily be made in any ordinary IC process. For this reason there again may be a preference for current sensing and thus for current as the information-carrying quantity.

2.6 Parasitics

Let us now, as fourth step in our design considerations, address the influence of parasitic immitances. The influence of parasitic admittances in parallel with the signal path can be reduced by terminating the signal path with a low impedance. The parasitic admittances then have no voltages across their terminals and thus no current flows in them. The influence of parasitic impedances in series with the signal path can be reduced by terminating the signal path with a high impedance. Then no current flows in the parasitic impedances and thus there is no voltage across their terminals.

In low-power integrated circuits, often the parasitic admittances, i.e., the node capacitances, e.g., the transistor's junction capacitances, due to their (non-linear) voltage dependency, have more influence on the signal behavior than the parasitic impedances, i.e., the branch inductances and resistances, e.g., the transistor's bulk resistances. Therefore it is convenient to terminate the signal paths with low impedances as much as possible. In this situation it is best to choose current as the information-carrying quantity.

This argument is also at the base of the popularity of 'current-mode', 'switched-current' and 'switched-transconductance' techniques [2, 3, 4], of which it is rightly stated that they have an inherent ability to exhibit good high-frequency properties.

2.7 Power supply

Finally, we have to consider the power supply. In practice, this power supply is a voltage source (battery), giving a limitation in voltage. The limitation in current is only indirectly given by a limitation in the power of the battery and might be less restricting than that of the voltage. This favors the choice of current as the information-carrying quantity. However, not using the total range of this supply voltage for signal swing gives rise to waste of power [5]

2.8 Conclusions

From the above discussion it will be clear that for low-voltage low-power analog ICs the total design process must be considered, in which transfer quality plays a dominant role. The theory was verified in several practical applications, such as circuits for hearing instruments (see, e.g. [6]), of which an example, an automatic gain control, will be discussed in one of the next chapters. Our designs, based on the design principles shown above, confirm that current becomes more favorable than voltage as the information-carrying quantity and that indirect feedback is to be preferred in a low-voltage low-power environment.

References

[1] E.H. Nordholt, *Design of high-performance negative-feedback amplifiers*, Elsevier, Amsterdam, 1983.

[2] C. Toumazou, F.J. Lidgey and D.G. Haigh (editors), *Analogue IC design: the current-mode approach*, Peter Peregrinus, London, 1990.

[3] C. Toumazou, J.B. Hughes and N.C. Battersby (editors), *Switched currents: an analogue technique for digital technology*, Peter Peregrinus, London, 1993.

[4] C. Toumazou and N.C. Battersby, *Switched-transconductance techniques: a new approach for tuneable, precision analogue sampled-data signal processing*, in Proc. IEEE ISCAS, Chicago, May 1993.

[5] E.A. Vittoz, *Low-power design: ways to approach the limits*, in Proc. IEEE ISSCC, February 1994, pp. 14-18.

[6] A.C. van der Woerd and W.A. Serdijn, *Low-voltage low-power controllable preamplifier for electret microphones*, IEEE J. Solid-State Circuits, Vol. 28, No. 10, pp. 1052-1055, October 1993.

Chapter 3

Devices

Chris Verhoeven and Koert van der Lingen

3.1 Introduction

The intention of this chapter is to show the relation between some of the electrical parameters of devices and their layout. Two different groups of device parameters can be distinguished.

1. The *key parameters.*
   The key parameters are dominant in determining the behavior of the circuit that is implemented with the devices. The accuracy of their value determines the accuracy of the circuit. The numerical value of the key parameters is of importance. For example, in a band-gap reference, the parameter EG, which describes the band-gap energy at $0K$, is directly related to the output voltage of the band-gap reference.

2. The remaining parameters.
   These parameters do not influence the circuit behavior significantly when their value is above or below a certain value. For example, usually the current gain of a bipolar transistor used in a negative feedback amplifier just has to be larger than a certain minimal value to obtain accurate amplification. The accuracy of the current gain has no direct influence on the accuracy of the amplifier.

In a structured electronic design, the desired values of the key device parameters are usually known.
In some cases, the numerical value of a key parameter can be set by choosing an appropriate layout. In all cases, the accuracy of the parameters and the matching between the parameters of two devices of the same type can be controlled by the layout.

It is not the intention in this chapter to deal with device physics in detail. However, sometimes it will be necessary to discuss a part of the physical behavior of the devices in order to clarify the relation between the electrical device properties and the layout.

3.2 Currents in semiconductors

In this section some properties of semiconductor devices are discussed in an intuitive way. For a detailed survey of physical and electrical device behavior, there is much available literature [1, 2, 3]. This section is intended only to point out some properties that are important in layout design. In semiconductors there are two important mechanisms that can cause electric current. They are:

- diffusion
- drift

In most cases, both drift and diffusion currents are involved in the currents measured at the terminals of the devices. Before studying the currents in the devices, first we treat the mechanisms themselves.

3.2.1 Diffusion current

When in a PN junction P-type silicon and N-type silicon are brought into contact, a *diffusion* current flows between through the junction. The mechanism behind the diffusion current is *chance*. The electrons and holes in the silicon move around in an arbitrary way. The holes are not "aware" of a shortage of holes in the N-type silicon, and the electrons are not "aware" of the electron shortage in the P-type silicon. Fig.3.1 shows a one-dimensional junction.

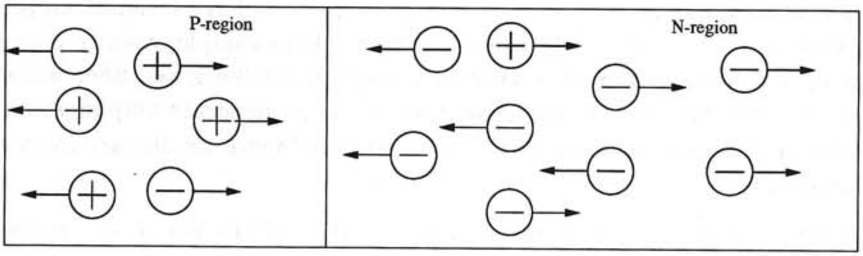


Figure 3.1: Sketch of a one-dimensional junction

In the P region there are four holes and two electrons, so it is P-type indeed. In the N region there are eight electrons and one hole. The holes and electrons have no preference for moving in a certain direction. Therefore, on the average, four of the electrons in the N region move to the left and the other four move to the right.

In the P region one of the electrons moves to the left and the other one moves to the right into the N region. Thus, effectively three electrons have moved from the right to the left. In a similar way, there is an effective flow of holes from the left to the right. This mechanism of carrier movements is called *diffusion current*. It is based on carrier density gradients and *not* on electric fields.

3.2.2 Saturated currents

Saturated currents are typical examples of currents that are not directly related to an electric field, but are governed by the diffusion mechanism. An example of this is the saturation current of a reverse-biased PN junction. In the P and N region, carriers, both majority and minority carriers, are thermally generated. The minority carriers diffuse to the depletion layer where they become subject to the electric field and cross the junction. This gives rise to a saturation flow at the junction. The electrons move to the N region and the holes to the P region. This movement of charge through the junction is a saturated current. The amount of charge that crosses the junction per second depends on the availability of carriers at the edges of the depletion layers and not on the magnitude of the electric field. Therefore, the voltage across the junction has no influence on the magnitude of the current, it merely provides the "means of transport" for every available carrier. Therefore, a junction biased in reverse behaves like a current source.

3.2.3 Drift current

When a charge carrier is put into an electric field, it undergoes a force in the direction of the field, which accelerates it in that direction. The acceleration is limited by collisions of the carrier to the lattice, which "reset" the velocity of the carrier again.

In a semiconductor, all charge carriers move randomly because of the thermal agitation. When an electric field exists, it can be imagined that this field "affects" the probability distribution of the direction of motion of each carrier. In the random movement, some "preference" for the direction indicated by the field can be observed. This results in an average movement in the direction of the field which is called *drift*.

The amount to which the carriers can respond to the force of the field is governed by the *mobility* μ.

3.3 The PN junction

The simplest device to introduce the current in a semiconductor is the PN junction, or the diode. The behavior of the diode for several different biasing conditions is treated below. After starting with the diode, we will later introduce the bipolar transistor.

3.3.1 The PN junction without external biasing

When the PN junction is in its thermal equilibrium, there is an electric field across the depletion layer. This implies there is a potential difference across the junction equal to the built-in voltage, V_{bi}. It can be seen as being just a contact voltage that is always generated when two different materials are brought into contact. Still, across the external terminals of a diode, this voltage cannot be measured. This is because in a physical diode three junctions are in series:

1. The actual PN junction.

2. A junction between the P-type silicon and the connecting metal.

3. A junction between the N-type silicon and the connecting metal.

Across the latter two junctions also the contact voltages, V_{pm} and V_{nm}, respectively, are found. The sum of these three voltage equals zero.

$$V_{bi} + V_{pm} + V_{nm} = 0 \tag{3.1}$$

No current flows from the P to the N region in steady state. The mechanism behind this can be understood as follows. When the P and the N region are brought into contact, a diffusion current of holes starts from the P to the N region simply because there are more holes in the P region than there are in the N region. Because of this, it is more likely that a hole will travel from P to N than in the other direction. Note that the holes are not aware of the gradient in their concentration. They are not "pushed out" of the P region due to overcrowding, they feel no "diffusion pressure". The diffusion current of electrons starts in a similar way.

Since the holes and the electrons both are charge carriers, their travel to the other region also implies the movement of charge to that other region. The electrons and the holes also leave an ion (fixed charge) behind. This results in a build-up of charge across the junction, and, consequently, a build-up of an electric field. This electric field causes the charge carriers to drift. The movement in the junction is no longer random, a certain average movement which is opposite to the direction of the diffusion current builds up. In the end, the field becomes strong enough to stop the diffusion current.[1]

[1] Of course this is a simple and handy interpretation of physical effects that are much more complicated in "reality", but for the average electronic designer, it suffices.

The energy bands, the space charge ρ, the electric field ϵ and the potential ψ of a diode in thermal equilibrium are shown in Fig. 3.2. The energy levels and potentials are related by $\phi = -E/q$. The current density for the holes is:

$$J_\mathrm{p} = -q\mu_\mathrm{p} p \frac{\partial \psi}{\partial x} - qD_\mathrm{p} \frac{\partial p}{\partial x} \tag{3.2}$$

and for electrons it is:

$$J_\mathrm{n} = -q\mu_\mathrm{n} n \frac{\partial \psi}{\partial x} + qD_\mathrm{n} \frac{\partial n}{\partial x} \tag{3.3}$$

The first term in the previous two equations represents a drift flow, while the second term represents the flow of carriers by diffusion.

In thermal equilibrium the two terms are equal except for the sign, so the net results are equal to zero.

Assuming a doping concentration N_A in the P region and a concentration N_D in the N region the minority carrier concentrations are given by:

$$n_0 = \frac{n_\mathrm{i}^2}{N_\mathrm{A}} \quad \text{in the P region and} \tag{3.4}$$

$$p_0 = \frac{n_\mathrm{i}^2}{N_\mathrm{D}} \quad \text{in the N region} \tag{3.5}$$

3.3.2 Biasing the junction

When a voltage V is applied across the junction, the electric field across the junction is affected. It is enlarged when the junction becomes reverse biased, and it is reduced when the junction becomes forward biased. In the latter case, the diffusion currents are no longer completely counteracted by the field, so a diffusion current starts to flow. When a junction is forward biased, this is the dominant part of the current that is measured externally.

In Fig.3.2, the applied voltage changes the potential ψ by an amount $-V$. Thus, applying a voltage changes the first term of equations 3.2 and 3.3. The result is a flow of carriers across the junction. If the junction is forward biased, the drift term becomes negligible with regard to the diffusion term and a diffusion flow from the majority to the minority region results. This is called the *injection* of carriers [7].

The concentration of injected holes from the P region into the N region can be calculated [8]. At the border of the depletion region and the N region, the injected hole concentration is equal to:

$$p_\mathrm{excess} = p_0(\mathrm{e}^{qV/kT} - 1) \tag{3.6}$$

Here p_excess = the excess hole concentration at the N region depletion edge.

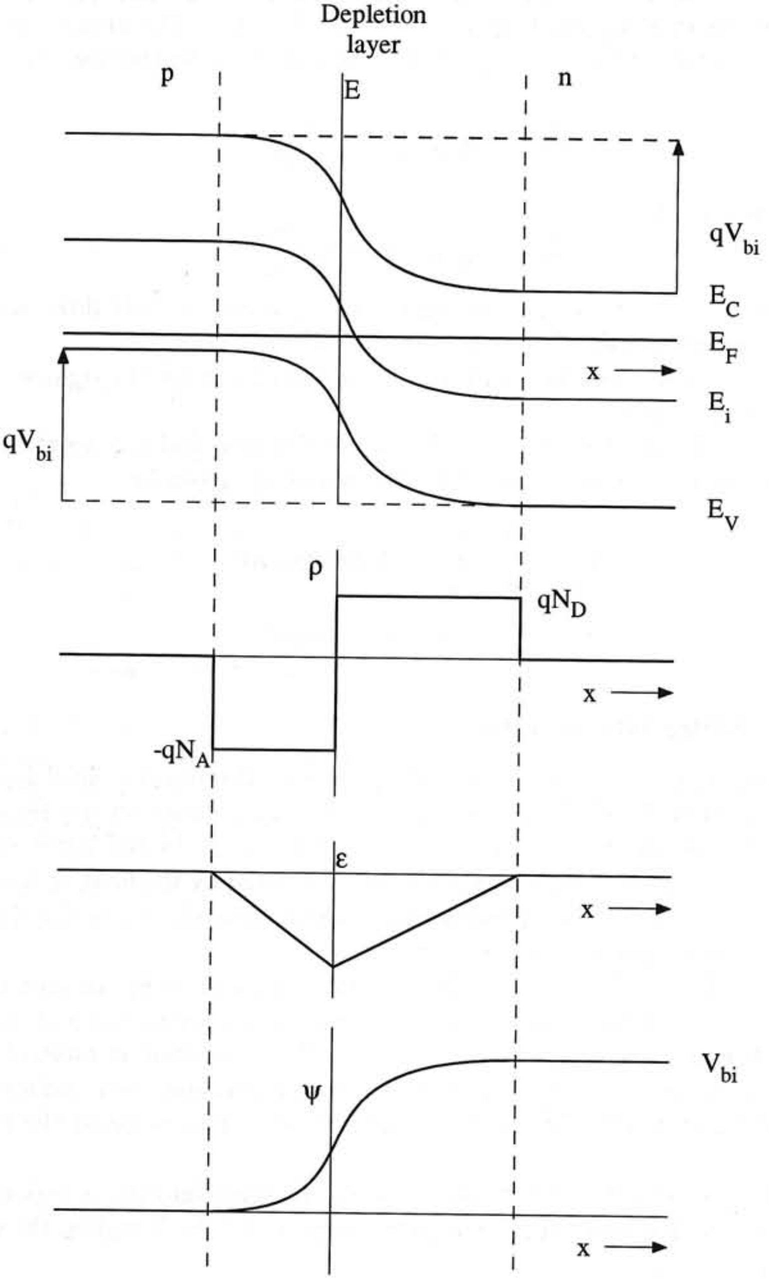


Figure 3.2: A PN junction in thermal equilibrium. From top to bottom: the energy bands, the space charge, the electric field and the potential.

p_0 = the hole concentration in the N region in thermal equilibrium as given in equation 3.5.

V = the voltage applied at the contacts.

k = Boltzmann's constant.

q = the electron charge.

T = the absolute temperature.

If the voltage is reversed, the drift term becomes much larger than the diffusion term and minority carriers are transported by the field from one side of the junction to the other side. This is called the *extraction* of carriers [7]. The transported carrier concentration can be found from equation 3.6. However, because the depletion layer is "empty", there are not many carriers for the field to transport. Only carriers that are generated thermally in the junction give rise to a saturated current as described in 3.2.2. This is described in the next section.

Similar equations can be derived for the electron concentration in the P region. A second-order effect of a biased junction is the modulation of the depletion width. Without going into detail, we can say that the depletion layer gets smaller when the junction is forward biased, and it gets wider when the junction is reverse biased.

Generation and recombination currents of a PN junction

Apart from transport of carriers across the junction, generation and recombination of carriers in the depletion layer occurs. The most important mechanism is the Shockley-Read-Hall mechanism. Here holes and electrons are generated thermally and recombine via traps, see for example [5, 6]. The carrier generation can be described by:

$$\frac{\partial}{\partial t} p_{\text{gen}} = \frac{\partial}{\partial t} n_{\text{gen}} = \frac{n_{\text{i}}}{\tau_{\text{eff}}} \tag{3.7}$$

The recombination rate is found to be:

$$\frac{\partial}{\partial t} p_{\text{rec}} = \frac{\partial}{\partial t} n_{\text{rec}} = \frac{n_{\text{i}}}{\tau_{\text{eff}}} \mathrm{e}^{qV/2kT} \tag{3.8}$$

where τ_{eff} is the effective lifetime of the carriers.

Carrier flows in a PN junction

In the previous sections, we showed that there can be four carrier flows in a biased PN junction.

1. A diffusion (injection) current of holes.

2. A diffusion (injection) current of electrons.

3. A flow of holes and electrons into the depletion region where they recombine.

4. A flow of holes and electrons that are generated thermally in the junction and driven apart by the electric field across the junction.

The diffusion flows are also known as injection flows, because excess minority carriers are injected into a region (from the side where they were a majority). The two other flows result from the generation and recombination of carriers in the depletion layer. In Fig. 3.3, the four carrier flows across a PN junction are shown.

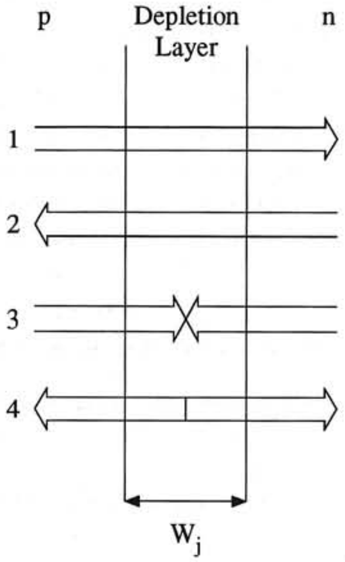


Figure 3.3: The carrier flows in a PN junction

3.3.3 Junction capacitance

The depletion layer in a reverse-biased PN junction is an "empty" space bounded by two areas in which there are charge carriers. This has much resemblance to a capacitor, and, indeed, a junction capacitance can be defined. The width of the depletion layer depends on the value of the reverse voltage that is applied, so the junction capacitance is voltage dependent. For an abrupt junction, the expression for the junction capacitance is:

$$C_j = A\left(\frac{q\epsilon}{2U_j\left(\frac{1}{N_A}+\frac{1}{N_D}\right)}\right)^{1/2} \tag{3.9}$$

in which N_A and N_D are the acceptor and donor concentration respectively, U_j is the voltage across the junction, q is the elementary charge, A the area of the junction and ϵ is the dielectric constant. The voltage across the junction is related to the voltage measured at the terminals U_{ext} via:

$$U_j = V_{bi} - U_{ext} \tag{3.10}$$

In most cases, the doping level of one of the regions is much higher than in the other. The depletion layer extends the most in the region that is has the lowest doping level. When this is N_A, (3.9) reduces to:

$$C_j = A\sqrt{\frac{q\epsilon N_A}{2(V_{bi} - U_{ext})}} = C_{j0}\frac{1}{\sqrt{1 - U_{ext}/V_{bi}}} \tag{3.11}$$

The right-hand expression is commonly employed to describe a junction capacitor, and C_{j0} and V_{bi} are parameters that are extracted via measurement. The square root relation only holds for an abrupt junction, which, in practice, it never is. This implies that the exact power of the non-linearity, which can be expected to deviate from the ideal factor 0.5, is also found by measurement and parameter extraction. Therefore, there are three parameters in the model for a junction capacitor. In SPICE they are called CJ0, the "zero-bias PN capacitance, VJ, the "PN potential" and M, the "PN-grading coefficient, which results in:

$$C_j = \mathrm{CJ0}\frac{1}{(1 - U_{ext}/\mathrm{VJ})^{\mathrm{M}}} \tag{3.12}$$

Of course, this relation only holds when the PN junction is in reverse bias. When it comes into forward bias, U_{ext} may become equal to or larger than V_{bi}. Still, it would be convenient to have a measure for the junction capacitance in this region, since there is still a relation between charge and voltage in a forward-biased junction which has the dimension of a capacitance. In SPICE, this problem has been solved by "switching" to another model when the limits of (3.12) are reached. An extra parameter FC is defined and when:

$$U_{ext} = \mathrm{FC} \cdot \mathrm{VJ} \tag{3.13}$$

instead of (3.12) the following expression is used:

$$C_j = \mathrm{CJ0}(1 - \mathrm{FC})^{-(1+\mathrm{M})}\left(1 - \mathrm{FC}(1 + \mathrm{M}) + \mathrm{M}\frac{U_{ext}}{\mathrm{VJ}}\right) \tag{3.14}$$

Generally, all parameters are found by measurement and curve-fitting.

Non-linearity

The voltage dependence of the junction capacitance causes non-linearity when it is used in a signal path. An example of this can be seen in Fig.3.4. This figure shows the result of a simulation in which two capacitors are charged with a constant current. One capacitor is a normal capacitor, the other is a junction capacitor. It can be seen that when the voltage across the capacitors rises, the value of the junction capacitor reduces. This reduction of the capacitance results in a faster change of the voltage at a constant value of the charging current. In Fig.3.5 the result can be seen of a simulation in which a sinusoidal voltage is applies across the diode. It has been given a DC offset to keep the diode reverse biased. At large reverse voltages, the capacitor is the smallest, and, consequently, the current is the smallest. At small junction voltages, the capacitance is the largest, and,

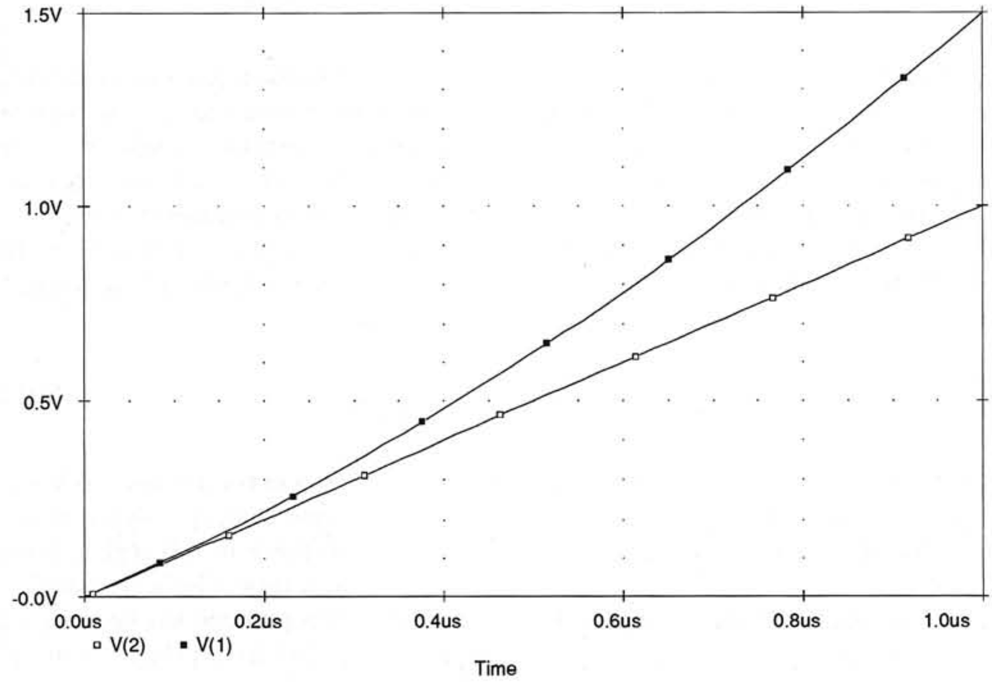


Figure 3.4: Voltage across a normal capacitor and a junction capacitor, both charged with an equal constant current

therefore, the currents are also. It can be clearly seen that the current through the junction capacitor has become distorted by this phenomenon. From all this, it can be concluded that when the linearity of a circuit is critical, it is very wise during the design to model capacitors that are implemented with junction capacitors by reverse-biased diodes.

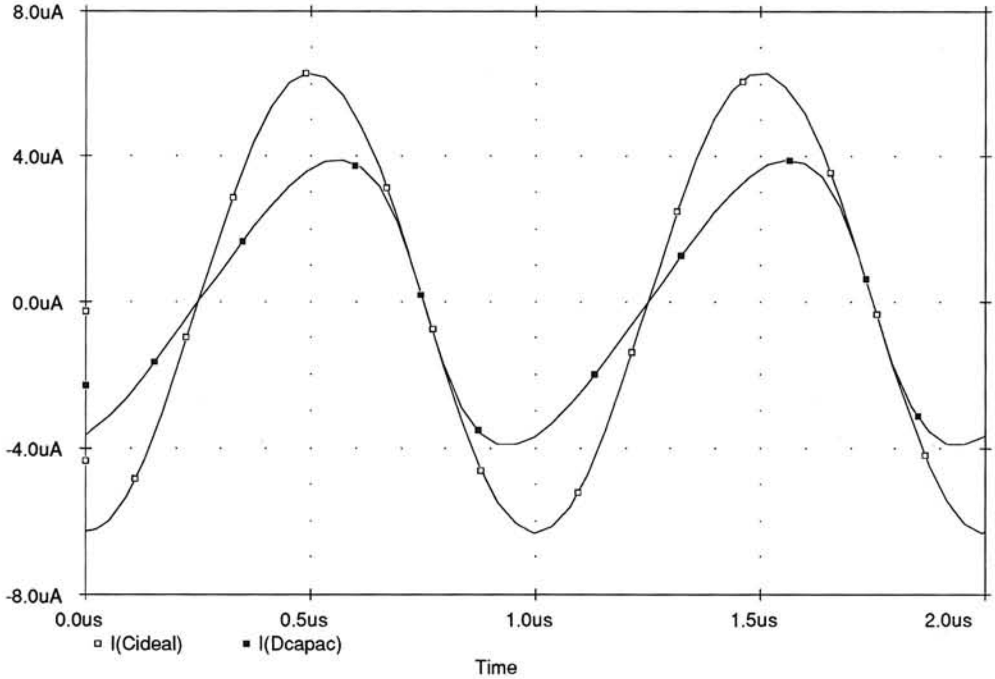


Figure 3.5: The current through a normal capacitor and a junction capacitor when excited with a sinusoidal voltage

True capacitors It is also possible to use "real" capacitors, consisting of two conducting plates, with an insulator in between. The capacitance between two metal interconnect layers may be used, or the capacitance between a poly layer and a metal layer. In some cases, there are special processing steps available to enable a thinner insulator between the capacitor plates than is standard for the two layers. At present, these capacitors offer less capacitance per unit area than junction capacitors, so a price is paid for linearity. Another problem that applies to all types of integrated capacitors is the fact that they are never truly floating. Always, at least one of the plates has a parasitic capacitance to the substrate. The value of this parasite may be up to a quarter of the value of the integrated capacitor. This value is of such significance that it is good practice to include it in the design from the start.

3.4 The bipolar transistor

In this section, we discuss the behavior of the bipolar transistor. During the discussion, the modeling of the transistor in SPICE is kept in mind. In all places where parameters occur that are also present in SPICE, the SPICE names are used in the equations.

3.4.1 Carrier flows in the bipolar transistor

The carrier flows through a junction depend on the biasing of the junction. In a transistor, biased in the forward active mode, some flows can be neglected in the different junctions. A schematic of a bipolar transistor is drawn in Fig. 3.6. The emitter, base and collector parts are designated. The carrier flows through the junctions are also shown. We assume an NPN transistor.

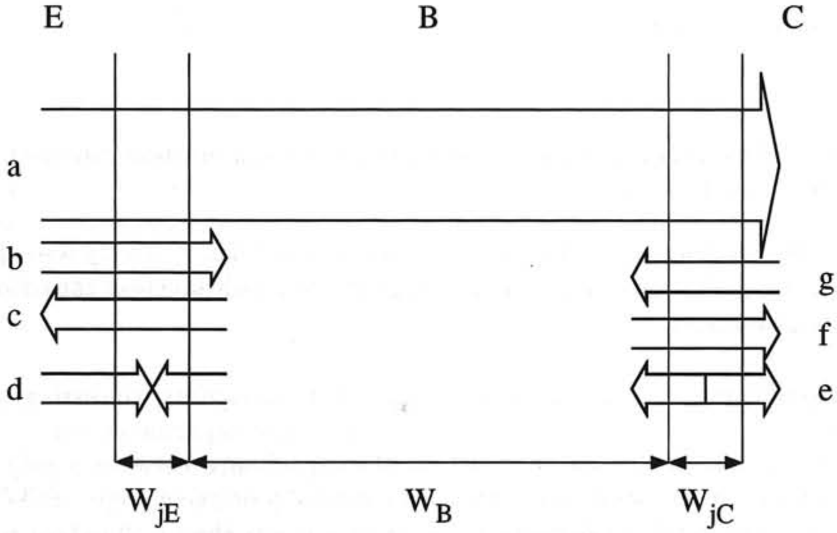


Figure 3.6: Schematics of carrier flow in a bipolar transistor

a: The flow from emitter to collector

The ideal transistor action is the injection of electrons from the emitter into the base, where these electrons diffuse to the base-collector junction. Because of the electric field across this junction, the electrons are swept through the depletion layer to the collector. This flow forms the main part of the collector current. The concentration of electrons arriving at the collector is given by:

$$n_{\mathrm{excess}} = n_0(\mathrm{e}^{qU_{be}/kT} - 1) \tag{3.15}$$

This flow is the main part of flow no.1 in Fig. 3.3.

b: Carriers injected from the emitter recombining in the base

A minor part of the electrons injected from the emitter recombine in the base with holes. This is the remainder of flow no.1 of Fig. 3.3. The recombining carriers add to the emitter and base current. The recombining concentrations are usually negligible.

c: Carriers injected from the base into the emitter

Because of the forward biasing of the emitter-base junction, holes are also injected from the base into the emitter. The injected carriers diffuse through the emitter and thus contribute to the base and emitter current, but not to the collector current. The same calculation as for the electrons can be made for the hole flow, which results in (3.6). In Fig. 3.3, this flow is designated as no.2.

d: Carriers recombining in the emitter-base depletion region

In a forward-biased junction, recombination of carriers in the depletion layer dominates the generation. This is because there are many charge carriers traveling through the forward-biased junction, so the chance of recombination is increased. In Fig. 3.3 this is flow no.3. The recombination process has already been described. The recombination rate is given by equation 3.8. The carrier flows affect the base and the emitter currents.

e: Generated carriers in the base-collector junction

The base-collector junction is either reverse biased or the junction voltage is zero. With a zero voltage, the recombination and generation in the depletion layer are in equilibrium. In a reverse-biased junction, there is a net flow of generated carriers out of the depletion layer, flow no.4 in Fig. 3.3. The generated carriers contribute to the base and the collector currents.

f,g: The saturation flow of the base-collector junction

In the base and collector regions, carriers are thermally generated. These carriers diffuse to the depletion layer, where they become subject to the electric field and cross the junction. This gives rise to a saturation flow at the base-collector junction. In Fig. 3.3, the saturation flow is formed by flows no.1 and no.2.

3.4.2 Operating regions of the bipolar transistor

Operation of the transistor can be divided into three regions. These regions are determined by the magnitude of the injected carrier flow (flow a). These regions are sketched in Fig. 3.7. In these three regions, it is assumed that in thermal equilibrium the minority carrier concentration in the base is negligible with regard to the base doping concentration.

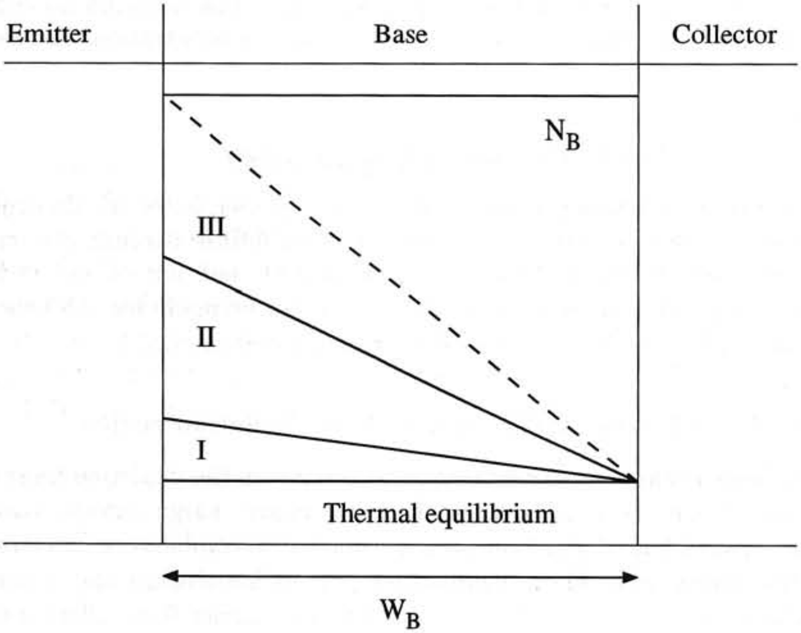


Figure 3.7: The operating regions of the transistor. The charge concentrations in the base for each operating region are shown. The base doping level N_{B} is also shown. Not to scale.

Operating region I

In region I, the concentration of injected electrons from the emitter is in the order of the minority electron concentration n_0 in the base. This is low-level injection.

The excess electrons transported to the collector are described by equation 3.15. The mentioned non-ideal carrier flows must also be taken into account.

This region is of particular interest in low-voltage low-power design, since it are the non-idealities that occur in this region that set the lower limit for the currents and voltages at which the transistors can still be used.

Operating region II

In region II, moderate-level injection begins, when the electrons, injected into the base, outnumber the electrons previously present. Neglecting the latter, the injected electron concentration can be written as:

$$n_{\mathbf{excess}} = n_0 \mathrm{e}^{qU_{be}/kT} \tag{3.16}$$

All non-ideal carrier flows can be neglected with regard to the injected electron flow. This is the "ideal" operating region, because the collector current is formed by the ideal transistor action.

Operating region III

Region III starts when the injected electron concentration is in the order of the base doping concentration. This is known as high-level injection. Because charge neutrality must be maintained, excess holes have to be supplied by the base contact. These excess holes can no longer be neglected in calculating the base minority carrier concentration. Therefore, equation 3.15 is no longer valid. The value at which the injected electron concentration equals the base doping level has special significance. This value has been indicated in Fig. 3.7 with a dotted line. It is known as the current at which high injection starts, and it is called the knee current.

For low-power design, it is never intended to have transistors operate at high-level injection. Therefore, not much attention is here paid to this effect. There are special cases, for example, at very high temperatures, where the effect becomes inevitable, but these circumstances are beyond the scope of this book.

3.4.3 DC currents in the bipolar transistor

The carrier flows previously discussed, transport charge from one place to another. The amount of charge transported per unit of time is a current.

The current through a junction is ideally determined by the injected minority carriers. The resulting current can be written as the amount of injected carriers that are transported times the charge per carrier q, times the diffusion constant, times a factor accounting for geometrical aspects such as area and diffusion length. The total current is the sum of the hole current and the electron current.

In a bipolar transistor, the situation is somewhat different. The ideal collector current is formed by carriers, which are injected from the emitter, diffuse through the base and reach the collector. This is flow a in Fig. 3.6. The carriers injected from the base into the emitter, flow c, do not contribute to the collector current I_c. We can thus write:

$$I_c = qD_B F_{geo} n_{\text{excess}} \tag{3.17}$$

This equation is valid for NPN transistors. Factor F_{geo} represents the influence of the area and the width of the base.

The currents in region II

In the case of moderate injection, the injected concentration is well below the base doping level, so the majority carrier concentration equals the base doping concentration N_B. At the same time, the present minority carrier concentration can be neglected with respect to the injected carrier concentration. According to equation 3.5, the minority electron concentration in the base is:

$$n_0 = \frac{n_\text{i}^2}{N_\text{B}} \tag{3.18}$$

The injected electrons from the emitter are described by (3.16). Substituting this in the equation for the collector current results in the following:

$$I_c = A \frac{qn_i^2}{\int_0^{\text{Wb}} \frac{N_B(x)}{D_B(x)} dx} e^{qU_{be}/kT} \tag{3.19}$$

A is the area of the base and Wb is the distance from the collector junction to the emitter junction, called the base width. The integral of N_B and D_B is taken because, generally, both parameters tend to vary across the base region. The integral is know as the *Gummel number* of the base. With:

$$\text{IS} = A \frac{qn_i^2}{\int_0^{\text{Wb}} \frac{N_B(x)}{D_B(x)} dx} \tag{3.20}$$

the collector current is written as:

$$I_c = \text{IS} e^{qU_{be}/kT} \tag{3.21}$$

This is the ideal relation between the emitter-base voltage and the collector current.

Adjacent to the electron flow is a hole flow from base to emitter. Similar to the collector current, it can be written as:

$$I_{b,inj} = qD_E F_{geo,E}\, p_{excess} \tag{3.22}$$

$F_{geo,E}$ is the factor that accounts for the emitter geometry and D_E is the diffusion coefficient for holes in the emitter part.

The current associated with the hole flow can be described as:

$$I_{inj} = \frac{\mathrm{IS}}{\mathrm{BF}} e^{qU_{be}/kT} \tag{3.23}$$

BF is the forward current gain of the transistor and is defined as

$$\mathrm{BF} = \frac{\int_0^{Le} \frac{N_E(x)}{D_E(x)} dx}{\int_0^{\mathrm{Wb}} \frac{N_B(x)}{D_B(x)} dx} \tag{3.24}$$

where N_E is the emitter doping concentration.

The Gummel number of the emitter is found in the nominator. Currently, the distance between the emitter-base junction and the contact to the emitter (We) is usually larger than the diffusion length Le of the holes. For this reason, Le is found in the integral instead of We. However, in modern IC processes, such as processes with poly-silicon emitters, the emitters' widths become smaller and smaller, which finally causes We to reappear in the expression.

It can be seen from the expression that, for a high current gain, the following constraints are imposed:

- A small base width
- A high emitter doping
- A low base doping
- A large enough emitter width

The latter demand becomes important when the vertical dimensions of the transistor are shrunk so much that We reappears in the expression.

Equation (3.24) is, of course, only valid when recombination and generation in the base play a negligible role. For example, in region I this is not true, and, consequently, there the current gain is reduced.

The currents in region I

Lowering the emitter-base voltage results in a lower injection of charge carriers from the emitter. When the injection has become so low that the minority carrier concentration can no longer be ignored, the low-level injection region is entered. The transport current from the emitter to the collector must be described by equation 3.15. This gives

$$I_{c,inj} = \mathrm{IS}(e^{qU_{be}/kT} - 1) \tag{3.25}$$

$$I_{b,inj} = \frac{\mathrm{IS}}{\mathrm{BF}}(e^{qU_{be}/kT} - 1) \tag{3.26}$$

Further, the non-ideal carrier flows d, e, f and g in Fig. 3.6 have to be accounted for. Because these carriers do not diffuse through the base, equation 3.17 cannot be used.

The recombining carriers in the emitter-base junction, flow d, give rise to a recombination current $I_{b,rec}$, which can be written as:

$$I_{b,rec} = \mathrm{ISE}e^{qU_{be}/2kT} \tag{3.27}$$

where

$$\mathrm{ISE} = qA_e\mathrm{Wje}\frac{n_i}{\tau} \tag{3.28}$$

in which Wje is the width of the E-B depletion layer and A_e the area of the junction. In the base-collector junction, three non-ideal carrier flows are present. Because the junction is never forward biased, the resulting currents are the saturation current and the generation current. The generation current is:

$$I_{gen} = A_c q\mathrm{Wjc}\frac{n_i}{\tau} \tag{3.29}$$

in which Wjc is the width of the C-B depletion layer and A_c the area of the junction. The generation current is in a first approximation independent of the junction voltage, but if the junction voltage is zero, then the current is balanced by the recombination current I_{rec}.

The saturation flows f and g across the base-collector junction result in saturation currents. For these, we can write similar equations as for the emitter-base junction. A reverse current gain factor BR can be defined in a similar way as BF.

The total base and collector currents amount to:

$$I_b = I_{b,inj} + I_{b,rec} - I_{b,sat} - I_{gen} + I_{\mathrm{rec}} \tag{3.30}$$

$$I_c = I_{c,inj} + I_{c,sat} + I_{b,sat} + I_{gen} - I_{\mathrm{rec}} \tag{3.31}$$

The currents in region III

In operating region III, the injected minority carrier concentration is in the order of the base doping concentration. Therefore, it is permissible to neglect the non-ideal flows d to g of Fig. 3.6. A special current IKF is defined. At this current the electron concentration equals the base doping concentration N_B. A similar parameter IKR models the high injection effects when the transistor is biased in the reverse mode.

The effects are included by way of a multiplication factor used to scale the ideal collector current as:

$$I_c = \frac{I_{ideal}}{\frac{1}{2}\left(1 + \sqrt{1 + 4\left[\frac{I_c}{\mathrm{IKF}} + \frac{I_c(\mathrm{reverse})}{\mathrm{IKR}}\right]}\right)} \tag{3.32}$$

3.4.4 Early effect

In the previous sections, the device physics of the bipolar transistor were examined and currents were linked to the charge flows. No attention has yet been paid to a second-order effect which originates from changing junction voltages. Because the depletion layer width changes with the voltage across the junction, in a bipolar transistor, the base width changes with changing emitter-base and changing base-collector voltages. This in turn changes the collector current.

Similarly to the introduction of the high injection effects, the Early effect is also accounted for via a multiplication factor. The variation of the width of the BC depletion layer is modeled via a parameter VAF, known as the forward Early voltage. The variation of the width of the BE depletion layer is modeled via a parameter VAR, the "reverse" Early voltage. *Although the latter parameter has the word "reverse" in its name, it models equally well the variation of the width of the BE depletion layer in forward mode too!* Because the doping of the emitter is much larger than that of the base, it can even be expected that VAR is even lower than VAF. Its neglect can cause serious errors as is, for example, shown later in this book for band-gap references. VAR is usually measured with the transistor biased in reverse mode. In some cases, rightful doubt may exist if the parameter extracted in this way is accurate enough for application in a forward-mode situation.

The influence of the Early effect is modeled in the following way:

$$I_c = I_{ideal}\left(\frac{1}{1-\frac{U_{be}}{\mathrm{VAR}}+\frac{U_{bc}}{\mathrm{VAF}}}\right) \tag{3.33}$$

3.4.5 The SPICE model

The model used in SPICE, the Gummel and Poon model, which describes the current-voltage relations, is discussed in this section. It is not our intention to make a detailed study of the model, but to relate some of the parameters used in the model to the physical effects discussed above. This is necessary to relate these parameters to the layout of the transistor.

Apart from the DC behavior described above, the AC behavior insofar as it is of importance in low-power circuits is also discussed.

The circuit diagram according to the Gummel and Poon model is depicted in Fig. 3.8. The relations between the junction voltages and the collector and base currents, which originate from injection, only are modeled with two ideal diodes and one current source. The two diodes account for the base currents I_{bc1}/BR and I_{be1}/BF, while the current source $(I_{be1} - I_{bc1})/K_{qb}$ accounts for the collector current. Factor K_{qb} models the Early and the high injection effects. It contains parameters IKF, IKR, VAF and VAR as described in the previous section.

BF and BR are the forward and the reverse current gain of the transistor.

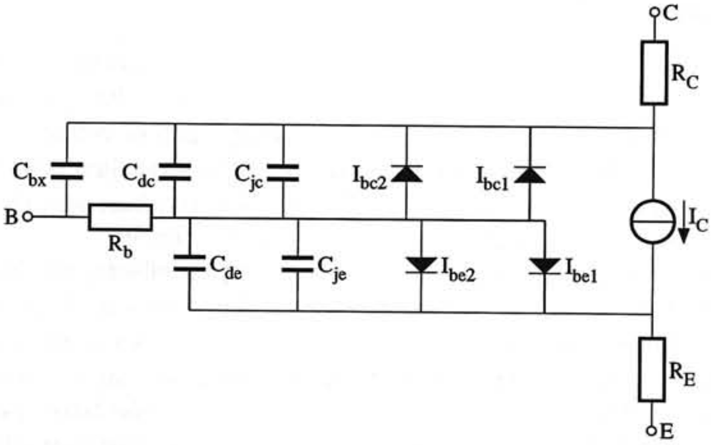


Figure 3.8: The circuit diagram of the Gummel and Poon model

Currents I_{bc2} and I_{be2} represent recombination and generation in the depletion layers.

AC behavior

Several capacitors are present in the model. Some of them are junction capacitors as described in section 3.3.3. The others represent the charge that flows through the junction and are known as *diffusion* capacitors.

The junction capacitance C_{jc} is distributed across the base resistance R_{B} by addition of the capacitor C_{bx}. The model includes the effect of high-level injection and base widening, by empirically modifying the equation for the collector current.

The charge-storage effects are modeled with four capacitors, C_{je} and C_{jc}, both being junction capacitors, for the charge storage in the junctions, and C_{de} and C_{dc} for the diffusion of charge carriers through the junctions.

The expressions for the junction capacitors are similar to those of the diodes. For the base-emitter junction the expression is:

$$C_{je} = \mathrm{CJE}\frac{1}{(1 - U_{be}/\mathrm{VJE})^{\mathrm{MJE}}} \tag{3.34}$$

For the base-collector junction, there is a similar expression with parameters CJC, MJC and VJC. The distribution of C_{jc} across the base is governed by parameter XCJC according to:

$$C_{jc} \;=\; \mathrm{XCJC}\;.\;\mathrm{CJC}\frac{1}{(1 - U_{bc}/\mathrm{VJC})^{\mathrm{MJC}}} \tag{3.35}$$

$$C_{bx} = (1 - \mathrm{XCJC}) \, . \, \mathrm{CJC}\frac{1}{(1 - U_{bc}/\mathrm{VJC})^{\mathrm{MJC}}} \tag{3.36}$$

$$\tag{3.37}$$

The diffusion capacitance, if present, is not affected by XCJC and does not contribute to C_{bx}.

As was the case for the PN junction, also here the model is switched when the junction becomes forward biased. For:

$$U_{be} = \mathrm{FC.VJE} \tag{3.38}$$

The expression for the junction capacitance becomes:

$$C_{je} = \mathrm{CJE}(1 - \mathrm{FC})^{-(1+\mathrm{MJE})}\left(1 - \mathrm{FC}(1 + \mathrm{MJE}) + \mathrm{MJE}\frac{U_{be}}{\mathrm{VJE}}\right) \tag{3.39}$$

A similar equation is found for the base-collector junction. Parameter FC is identical.

The diffusion capacitance is modeled via parameter TF for the base-emitter junction and via TR for the base-collector junction. Under normal bias conditions, only the base-emitter diffusion capacitance is of importance. The value is:

$$C_{de} = \mathrm{TF}\frac{I_c}{kT/q} \tag{3.40}$$

Parameter TF is approximately proportional to the base width, so to obtain an HF transistor, the base should be as thin as possible.

The total capacitance at the base-emitter junction is now:

$$C_{de} = \mathrm{TF}\frac{I_c}{kT/q} + \mathrm{CJE}\frac{1}{(1 - U_{be}/\mathrm{VJE})^{\mathrm{MJE}}} \tag{3.41}$$

For small signal analysis, the Taylor expansion of the expression around the bias point can be used:

$$C_{de} = \mathrm{TF}g_m + \mathrm{CJE}\frac{1}{(1 - U_{be}/\mathrm{VJE})^{\mathrm{MJE}}} \tag{3.42}$$

in which g_m is the transconductance of the transistor.

The emission coefficients

In the equations for the various currents discussed so far, an exponential factor has appeared of the type $e^{qU/kT}$ or $e^{qU/2kT}$, depending on the origin of the current.

In SPICE the factors 1 or 2 before kT appears as parameters. For the injection currents, the expressions become:

$$I_{be1} = \mathrm{IS}^{qU_{be}/\mathrm{NF}kT} \tag{3.43}$$

a similar expression is found for I_{bc1}, with parameter NR. According to the strict physics, these parameters should be exactly equal to 1. Accurate measurements usually yield values close to one, and since the parameter is multiplied by the temperature T, deviations from unity are easily explained as arising from errors in the measurement of the exact temperature of the device during measurement. Usually, this error is bigger than the deviation of the parameters from unity.

The non-ideal currents that should have a factor of two in the exponent are now written as:

$$I_{be2} = \mathrm{ISE}e^{qU_{be}/\mathrm{NE}kT} \tag{3.44}$$

a similar expression is found for I_{bc2}, with parameter NC. A parameter value lower than 2 is frequently found. This can be explained by the fact that a practical transistor is a three-dimensional device that can be seen as a parallel connection of several transistors of a different geometry and with different parameters. The average parameter can, therefore, lie between the physically possible values 1 and 2.[2]

The SPICE parameters

When a transistor is biased in the forward mode, a limited set of parameters has to be known in order to model the behavior properly. They are:

- IS
  The saturation current.

- NF
  The forward current emission coefficient. It should be equal to unity.

- BF
  The forward current gain.

- ISE
  The base-emitter leakage saturation current.

- NE
  The base-emitter leakage emission coefficient.

[2] There are also mechanisms that cause a factor of 4. However, in modern transistors, these mechanisms are hardly of any significance.

- VAF
  The forward Early voltage.

- VAR (!)
  The reverse Early voltage, also of importance in the forward mode.

- IKF
  The current at which high injection starts. In a lot of cases, the exact value is of no importance, since for proper circuit design it is better to avoid this region. However, it has to be known accurately enough to make sure high injection will not occur in a circuit.

- TF
  The ideal forward transit time.

- CJE
  The junction capacitance of the base-emitter junction at zero bias voltage.

- VJE
  The base-emitter built-in potential.

- MJE
  The base-emitter PN-grading factor.

3.4.6 The f_T

The f_T of a transistor is the frequency at which the current gain becomes equal to unity. It is measured as depicted in Fig.3.9a. The transistor is driven by an ideal current and it is short-circuited at its output.

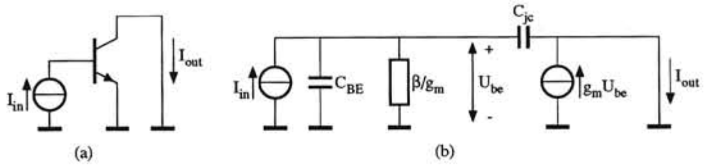


Figure 3.9: (a) Circuit for f_T measurement, (b) the small signal equivalent

From Fig.3.9b, it can be seen that the f_T of the transistor is:

$$f_T = \frac{g_m}{2\pi(C_{BE} + C_{jc})} \tag{3.45}$$

Capacitor C_{BE} contains contributions of the diffusion and the junction capacitor as given in (3.41).

For large currents, the diffusion capacitor becomes dominant and the expression for C_{BE} reduces to:

$$C_{be} = g_m \mathrm{TF} \tag{3.46}$$

For the f_T it follows then:

$$f_T = \frac{g_m}{2\pi(g_m \mathrm{TF} + C_{jc})} = \frac{1}{2\pi \mathrm{TF} + 2\pi C_{jc}/g_m} \approx \frac{1}{2\pi \mathrm{TF}} \tag{3.47}$$

This is the maximum f_T for the transistor, and it is independent of the current when the current is sufficiently high.
At small currents, the diffusion capacitor becomes negligible and the equation for f_T yields to:

$$f_T = \frac{g_m}{2\pi(C_{je} + C_{jc})} = I_c \frac{q/kT}{2\pi(C_{je} + C_{jc})} \tag{3.48}$$

It can be seen that now f_T has become current dependent and decreases with decreasing current. This is because of the fact that g_m reduces when I_c reduces whereas the junction capacitors are nearly independent of it. (The diffusion capacitor was proportional to the current, just like g_m.) This is an important effect when designing circuits for low currents. The f_T may be much lower than expected from parameter TF which is used to advertise $\frac{1}{2\pi \mathrm{TF}}$ as *the* transit frequency of the transistor.

3.4.7 Bulk resistors

There are three resistors shown in Fig.3.8. Two of them, RE and RC, are simple resistors formed by the material between the metal contact to the transistor and the actual junction. Both resistors are mainly important at high-current operation. The third resistor is the base resistance, which is dependent on the collector current. It is this resistor that causes most of the problems in circuit design, since it easily affects the noise behavior and the HF behavior of the transistor. Also it can give rise to emitter crowding, a phenomenon that causes the emitter current to concentrate into a limited part of the emitter. This effect degrades the performance of the transistor and may even damage it.

The base resistor

Noise caused by the base resistor The base resistor is in many cases one of the main causes of noise. In such a case, special layout techniques are used to reduce the base resistance. Finally, all layout techniques to reduce the base resistance come down to placing as many base resistances in parallel as possible. This can be achieved by either placing complete transistors in parallel or making as many base contacts as possible. In Fig.3.10, the latter technique is sketched. The

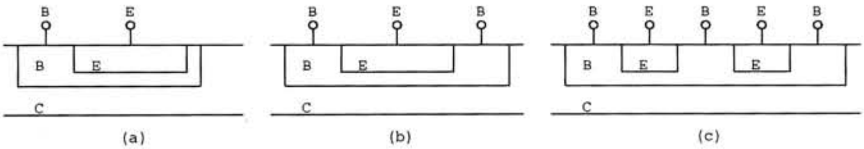


Figure 3.10: Different ways to implement base contacts

figure shows three typical integrated transistors in cross-section. It can be seen that the base width is smallest directly underneath the emitter. This is the "ideal" part of the transistor. The base current has to flow from the side to this part of the transistor, and since it is very thin, the resistance underneath the emitter can be expected to be relatively high.
For transistor (a), there is only a base contact at the right side, and the resistance from the contact to the left most point of the emitter is, therefore, formed by the complete path underneath the emitter. This, of course, provides a considerable contribution to the average base resistance of this transistor.
For transistor (b), the average resistance has already been much reduced, because of the extra contact on the left side of the transistor. Exactly in the middle, underneath the emitter, there is no lateral flow of the base current. The largest contribution to the average base resistance can, therefore, be only half the value of that seen in transistor (a).
Transistor (c) shows the most effective layout to reduce the base resistance. The emitters are made as thin as possible (as allowed by the design rules) in one direction. Thus, seen from the top, the transistor has a number of long thin "lines" as its emitter. Between those emitter lines, base contacts are placed. In the figure, it can be seen that the lateral distance for the base current in the high resistance area is thus made as short as possible.

Influence of the base resistance on the HF behavior The influence on the HF behavior of the transistor can best be understood from Fig.3.8. An AC input current for the transistor is divided between the path through R$_b$ and the path through C_{bx} and C_{jc}. (Note that in the correct f_T measurement, the collector is short circuited to the emitter, see Fig.3.9.) Normally, the latter current is negligible compared to the current flowing through C_{je}, C_{de} and the diodes. However, the current though C_{bx} becomes larger when the base resistance increases, thereby finally making its influence on f_T too large to neglect. Hence the f_T of the transistor is reduced.

In SPICE, C_{bx} is set to zero by default. Parameter XCJC is used to indicate what fraction of C_{jc} has to be assigned C_{bx}. To obtain accurate HF simulation, it should be checked that XCJC has an appropriate value. In this case also, the techniques shown in Fig.3.10 are useful to optimize the HF behavior. XCJC is strongly layout dependent.

Emitter crowding The base emitter part of a transistor is shown in Fig.3.11. The collector current flows from the metal emitter contact via the emitter bulk and then via the base to the collector. The base current flows from the metal base contact via a vertical path with resistance R_{b1} to the so-called intrinsic base region. There it flows from the left to the right through the intrinsic base. The resistance of the intrinsic base is indicated by the resistances $R_{b2\ldots n}$. The base current causes a voltage drop along the base region from the left to the right. This has as an effect that the local base to emitter voltage decreases from the left to the right. The reduced base emitter voltage results in a lower injection current. When the collector current is written as:

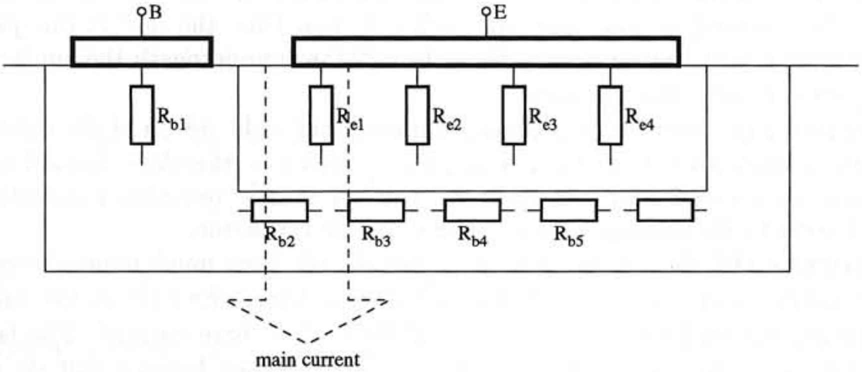


Figure 3.11: A sketch of the resistances and current flows in a transistor

$$I_c = \mathrm{IS}e^{qU_{be}/kT}, \tag{3.49}$$

it can be easily shown that to change the current by a factor n_{cr}, a change in the base emitter voltage is necessary, equal to:

$$\Delta U_{be} = \frac{kT}{q}\ln(n_{cr}) \tag{3.50}$$

From this it follows that to change the current by a factor of 2, the change in the base emitter voltage only has to be about 17mV. Since resistors $R_{b2\ldots n}$ can be easily in the order of kilo ohms, significant reductions occur at relatively small base currents! The "aggressive" reaction of the collector current to changes in the base emitter voltage is caused by the exponential relation between voltage and current.

When a transistor is being designed even for moderate currents, the placement of the base contacts still has to be considered with great care because of this phenomenon.

In Fig.3.11, it can be seen that due to the emitter crowding, the main current is flowing on the left side of the transistor. This implies that at the transistor's terminals, parameters are measured that are of this area. For the base resistance, this implies that only R_{b1} and R_{b2} play a significant role. The total value of the base resistance is therefore reduced when crowding occurs.

At very low currents, when there is no crowding, an average between the resistance of the shortest current path via R_{b1} and R_{b2} into the emitter and the longest path via R_{b1} and all resistors $R_{b2...n}$ into the emitter is found as the base resistance.

In SPICE, this reduction of the base resistance at high currents is modeled via three parameters:

- RB,
  the maximal base resistance (at zero base current).
- RBM,
  the minimal base resistance.
- IRB,
  the current at which the actual base resistance is between RB and RBM.

The parameters are usually found by measurement and curve fitting. The expression that is used is:

$$R_b = \frac{1}{A}\left(\mathrm{RBM} + 3(\mathrm{RB} - \mathrm{RBM})\frac{\tan(X) - X}{X\tan^2(X)}\right) \tag{3.51}$$

$$X = \frac{\sqrt{1 + \frac{144 I_b}{\pi^2 A\ \mathrm{IRB}}} - 1}{\frac{24}{\pi^2}\sqrt{\frac{I_b}{A\ \mathrm{IRB}}}} \tag{3.52}$$

in which A is the area of the transistor.

The default value for IRB is infinity. However, if it is not specified, a current dependency is still modeled in SPICE, because it switches to another model. Then the expression for the base resistance becomes:

$$R_b = \frac{1}{A}\left(\mathrm{RBM} + \frac{\mathrm{RB} - \mathrm{RBM}}{K_{qb}}\right) \tag{3.53}$$

Factor K_{qb} models the Early and the high injection effects. It contains parameters IKF, IKR, VAF and VAR as described before. Thus, setting IRB to a default makes the base resistance become dependent on completely different parameters. It is therefore important to know in which way parameters have been extracted and fitted, and also to use the supplied parameters as a complete set.

The emitter bulk resistor

RE reduces the transconductance of the transistor:

$$g_{m,eff} = g_{m,ideal}\frac{1}{1 + g_{m,ideal}\mathrm{RE}} \tag{3.54}$$

This effect is of special importance at high currents.

Though no current dependency of this resistor is modeled in SPICE, in practice it may show an increasing value for increasing currents. Due to emitter crowding, only a part of the transistor is active. This means that the current only flows through a part of the emitter. In this respect, all possible current paths in the emitter can be considered to be in parallel, so the emitter bulk resistor also consists of a great number parallel resistors. The effect is sketched in Fig.3.11. It can be seen that when the current is concentrated on the left side of this device, only R_{e1} plays a role, and that in this particular case the emitter bulk resistance is increased by a factor of about four.

The effect of the increase may partially compensate for the reduction of the base resistance, and might, therefore, "dissolve" in the behavior of the base resistance during parameter extraction. There are, however, rare cases in which the increase of the emitter bulk resistance is larger than the reduction of the base resistance. The obviously erroneous conclusion during parameter extraction may be that the base resistance increases at high currents.

The collector bulk resistor

The collector bulk resistor if formed by the bulk material between the active base-collector junction and external collector terminal. Since the collector has a low doping level, its value can be considerable. Usually, special measures are taken to reduce its value. The collector bulk resistance reduces the HF performance of a transistor and increases the risk of saturating the transistor.

Influence of RC on the HF behavior In section 3.4.6 the definition of the transit frequency f_T was given. It is measured by driving the transistor with a current and measuring the output current with the collector short circuited to the emitter. Due to the collector bulk resistor, the collector is no longer correctly short circuited to the emitter. Across the collector bulk resistor a voltage, which is in anti-phase with the base emitter voltage, is generated by the signal current. The voltage amplification factor is:

$$\mu = -g_m\mathrm{RC} \tag{3.55}$$

Across the junction capacitor between base and collector C_{jc}, there now exists a voltage that is increased by a factor $\mu + 1$. This implies that also the current through the capacitor is increased by this factor. As a result, the influence of C_{jc} is enlarged. It can be modeled as a capacitor with a value equal to:

$$C_{eff} = (\mu + 1)C_{jc} \tag{3.56}$$

connected in parallel with C_{BE}. The effect is known as the Miller effect.
At present, the value of a collector bulk resistor is in the order of tens to a few hundreds of ohms. This implies that even at modest currents, the apparent capacitor caused by C_{jc} can be increased by a factor of ten or more. In this case, the effective transit frequency can be reduced considerably, since the contribution of C_{jc} to f_T can no longer be neglected as it was for the derivation of (3.47). The f_T of the intrinsic transistor "cannot be reached" from the terminals due to the collector bulk resistor. Figure 3.12 shows a simulation result of an f_T measurement done for two different values of the collector bulk resistor. 0.1Ω has been taken as an ideal value, and 150Ω as a practical value. The latter value it not uncommon in standard bipolar processes. All other parameters are also taken from a standard bipolar process. It can be seen that in the ideal case, the current gain equals unity at about 4.8GHz, which is nearly equal to $\frac{1}{2\pi \mathrm{TF}} = 5.3\mathrm{GHz}$, the maximum f_T that could be achieved with this particular transistor. The mere introduction of a collector bulk resistor equal to 150Ω reduces the unity gain frequency to 4.0GHz, which is a reduction close to 1GHz.
From this it can be seen that the HF performance of an IC process cannot be evaluated just on the specification of f_T, and certainly not on TF. The other parameters may be such that this f_T is not available in practice.

Influence of RC **on saturation** When a transistor is in saturation, both its base-emitter and its base-collector junction are forward biased. Usually this is a situation to be avoided, since, for example, it makes the transistor slow and reduces its output impedance. Saturation can be prevented by keeping the voltage across the base-collector junction at least equal to zero. (In practice, it may even be chosen higher, if possible, to reduce the value of the junction capacitor C_{jc}.) A DC current (I_c) flowing through the collector bulk resistance, reduces the reverse bias voltage across the junction itself. When for an NPN transistor, the external bias voltage across the base-collector junction is kept equal to zero, the voltage at the junction is equal to:

$$U_{\mathrm{bc,internal}} = I_c \mathrm{RC}, \tag{3.57}$$

which is a forward biasing voltage. When saturation occurs, the output impedance of the transistor reduces to approximately the value of RC. The transistor starts behaving like a resistor instead of a current source.

In practice, zero bias voltages across the base-collector terminals of a transistor are because of this hazard, and certainty about the value of RC is necessary for low voltage applications for high frequencies, where the currents are considerable.

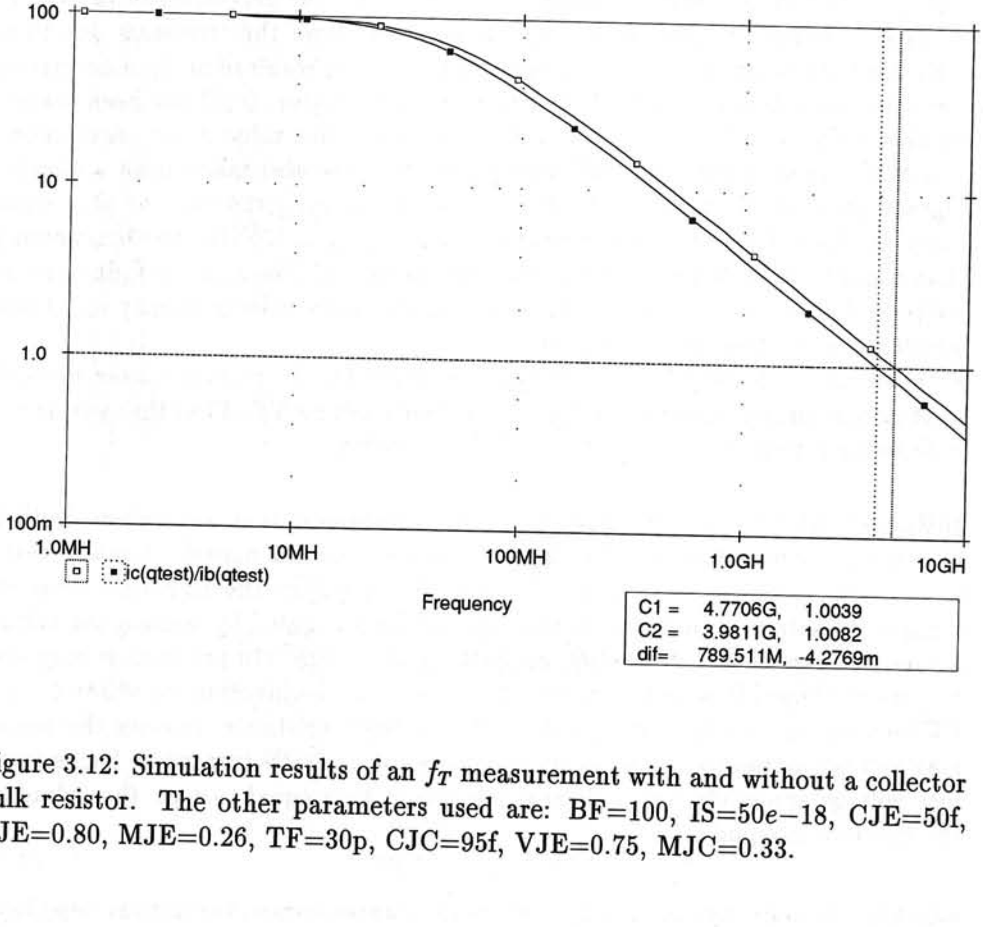


Figure 3.12: Simulation results of an f_T measurement with and without a collector bulk resistor. The other parameters used are: BF=100, IS=50$e-18$, CJE=50f, VJE=0.80, MJE=0.26, TF=30p, CJC=95f, VJE=0.75, MJC=0.33.

3.4.8 The layout of a bipolar transistor

Transistors in integrated circuits are embedded in the silicon bulk of the chip, separated from it by a reverse biased junction, an oxide barrier or a trench. In Fig.3.13 the layout of a simple bipolar transistor is shown. On a P-substrate, an N-type epitaxial layer is grown. A deep P implant (DP) is used to isolate an epi island. In the island, a base region is implanted, a lowly doped P region (BW) as the active (intrinsic) base region and a highly doped (extrinsic) P region (WP) for

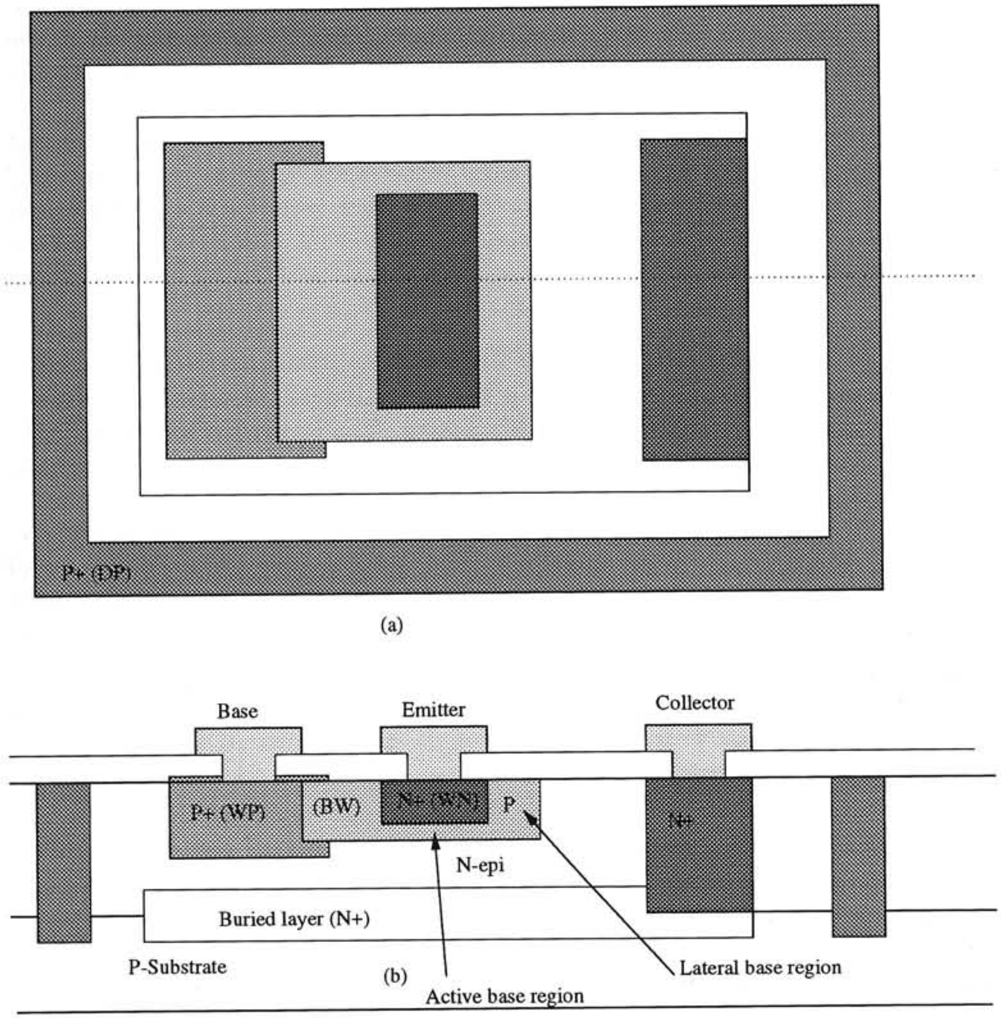


Figure 3.13: The layout of a typical bipolar transistor, (a) from the top, (b) a cross-section

connection of the active base region to external contact with a low resistivity path. The N-emitter (WN) is implanted in the BW region. To reduce the collector bulk resistance, a low resistivity N-type "collector plug" is implanted which connects to a highly doped N-type "buried layer". The buried layer is as close to the collector-base junction as possible. The distance depends on the voltages at which the transistor is supposed to operate. To keep the transistor performance optimal, the BC-depletion layer may never touch the buried layer,. A large amount of the collector bulk resistance finds it origin in the path from the buried layer to the BC-depletion layer.

In modern processes, there is much more sophistication than is shown here, but this simple sketch is sufficient to get an idea of the relations between the electrical behavior of the transistor and its layout. There is much variation between different IC-processes and the design manual of each process is the appropriate medium from which to obtain the details.

Parasitic PNPs The substrate turns the transistor into a four-terminal device. It is of the NPNP type, and looks like a thyristor. Indeed, thyristor operation is possible when the transistor is improperly biased. The layers of the base (P), the collector (N) and the substrate (P) form a PNP transistor. The performance of this transistor is intentionally degraded by giving it a high base doping. This is achieved with the buried layer. Apart from reducing the collector bulk resistance, its second (but actually most important) task is to reduce the current gain of the parasitic PNP. Even at low currents and frequencies, where the collector bulk resistance may play no role, the buried layer has to be there to degrade the PNP. When the collector base junction of the NPN becomes forward biased (saturation), this means that the emitter base junction of the parasitic PNP has become forward biased (same junction). If the PNP had enough gain, this would fire the thyristor and latch-up of the circuit occurs. Fortunately, because of the buried layer, the thyristors have not enough loop gain to stay on.

Of course, it is also possible to make intentional use of the PNPs. When the emitter implant of the NPN is left out and the buried layer is omitted, a PNP transistor is obtained of which the parameters are usually not too bad. A disadvantage is that the collector of these devices is inseparably connected to the global substrate node and therefore the collector current is not available as a separate current. For some applications, like medium performance band-gap references, the transistors can be of use.

There are some risks when signal currents are allowed to flow through the substrate, but this strongly depends on the layout and the specifications for the circuit.

When the parasitic PNPs are included in a SPICE netlist, it is possible to check for latch-up modes. If they are found, they are likely to occur in the practical circuit too and measures can be taken. If they are not found, unfortunately, this does not guarantee a correct functioning in practice.

The collector-substrate capacitor

From Fig.3.13, it can be seen that the collector contacts the substrate via, in this case, a reverse biased junction. Thus, a junction capacitor is found from the collector to the substrate. In SPICE, this capacitor is modeled with the parameters CJS, MJS and VJS, the equations are similar to (3.34).[3] This capacitor has in many cases a considerable influence on the HF behavior of the transistor, especially in combination with a collector bulk resistance. For Fig.3.14, the simulation of which the results were shown in Fig.3.12, is repeated with the addition of a collector-substrate capacitor, having a value of 0.25pF, again a realistic value for today. It can be seen that the combination of a collector bulk resistance and a collector-substrate capacitor results in an f_T of more than a gigahertz below the value expected from the "ideal" f_T.

Scaling and Matching

Looking at Fig.3.13, it can be seen that it is actually a three-dimensional structure, that can be modeled as the parallel connection of a vertical transistor and a lateral transistor. The vertical transistor is the "desired" transistor, the lateral transistor can be considered a parasite. The performance of the lateral transistor is worse in most respects. For example, it has a larger base width, so its current gain and its f_T are much lower. Also due to surface deficiencies, its low current properties are usually worse.

The behavior of the average transistor is measured at the terminal. The contribution of the vertical transistor to the total is proportional to the area of the transistor, and the contribution of the lateral transistor to the perimeter. From this, it follows that the area-sensitive transistor parameters do no scale exactly with the transistor area (as they are, for example, expected to do in SPICE). Also some parameters that are not sensitive to the area, like the Early voltages VAF and VAR, become layout dependent because the are sensitive to the *area-to-perimeter ration*, (SPR). Since the lateral and the vertical transistors have different parameters, and the SPR determines the relative magnitude of their contribution on the average, the SPR is an important layout aspect when matching is of importance. Two devices that should match do not only have to have equal areas, but also equal SPRs. Also when devices are scaled, not just the area, but also the SPR

[3] The (fourth) substrate terminal of a bipolar transistor is connected to ground (terminal 0) in SPICE by default. If this is not in accordance with reality, a fourth node number can be entered in the SPICE netlist for each transistor as the substrate node.

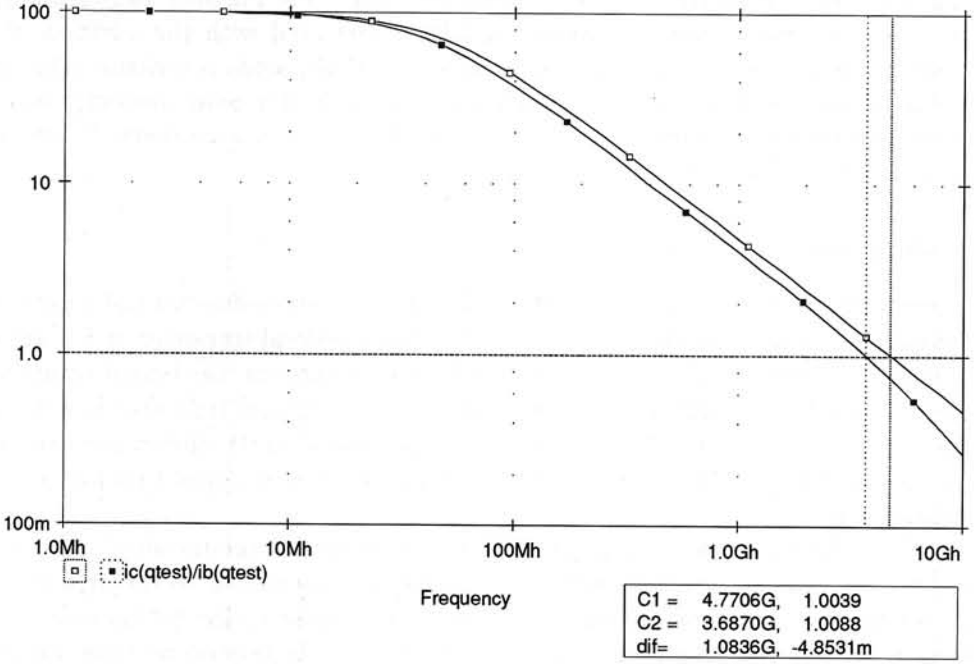


Figure 3.14: Simulation results of an f_T measurement with and without a collector bulk resistor and with a collector-substrate capacitor of 0.25pF. The other parameters used are: BF=100, IS=50$e-$18, CJE=50f, VJE=0.80, MJE=0.26, TF=30p, CJC=95f, VJE=0.75, MJC=0.33, MJS=0.33, VJS=0.75.

should be scaled with the same factor. Of course, exact scaling can be achieved when the factors are implemented by putting several "unit transistors" in parallel, but this is usually to harsh a measure. The SPR of the emitter base junction especially is important, so in most cases it suffices to put multiple unit-emitters in one large base region. To match the base resistance also, unit base-emitter regions can be put into one large collector region. Since the isolation layer around the transistors (DP layer), and the required spacing between this layer and the buried layer occupy a considerable space, this approach needs much less chip area than the parallel connection of complete transistors. Also in the latter case, the collector-substrate capacitor scales "correctly", which is usually not desired. This may be a serious argument against the optimal matching topology consisting of parallel connections of complete unit transistors. In Fig.3.15, various implementations for two transistors with a scaling 1:2 are shown. Transistor pair (a) has the worst matching, since the SPR has not been scaled properly. Pair (b) has matched emitters, pair (c) has matched base-emitter regions. It can be seen that pair (d),

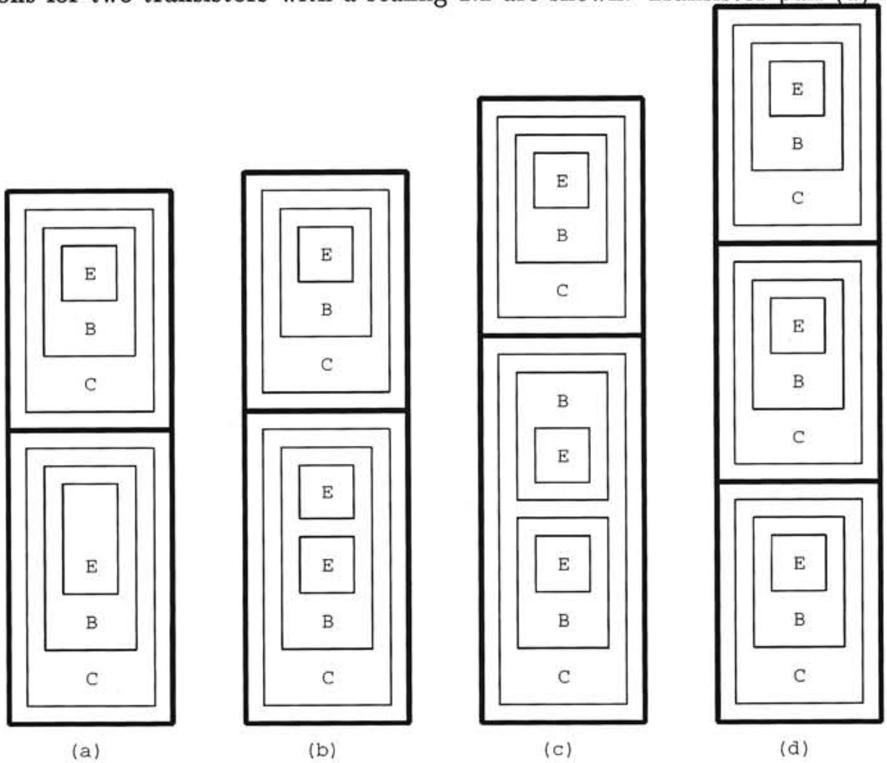


Figure 3.15: Different topologies to obtain a transistor pair that matches with a ratio 1:2

consisting of identical transistors placed in parallel, occupies most of the space.

3.4.9 The lateral PNP transistor

When a PNP transistor that has all three terminals available is needed, in many processes, the only option is the lateral PNP. In Fig.3.16, the layout of such a transistor is shown. Again it is a simple implementation, and in practice many variations are found. An N-type island is formed in a similar way as for the NPN.

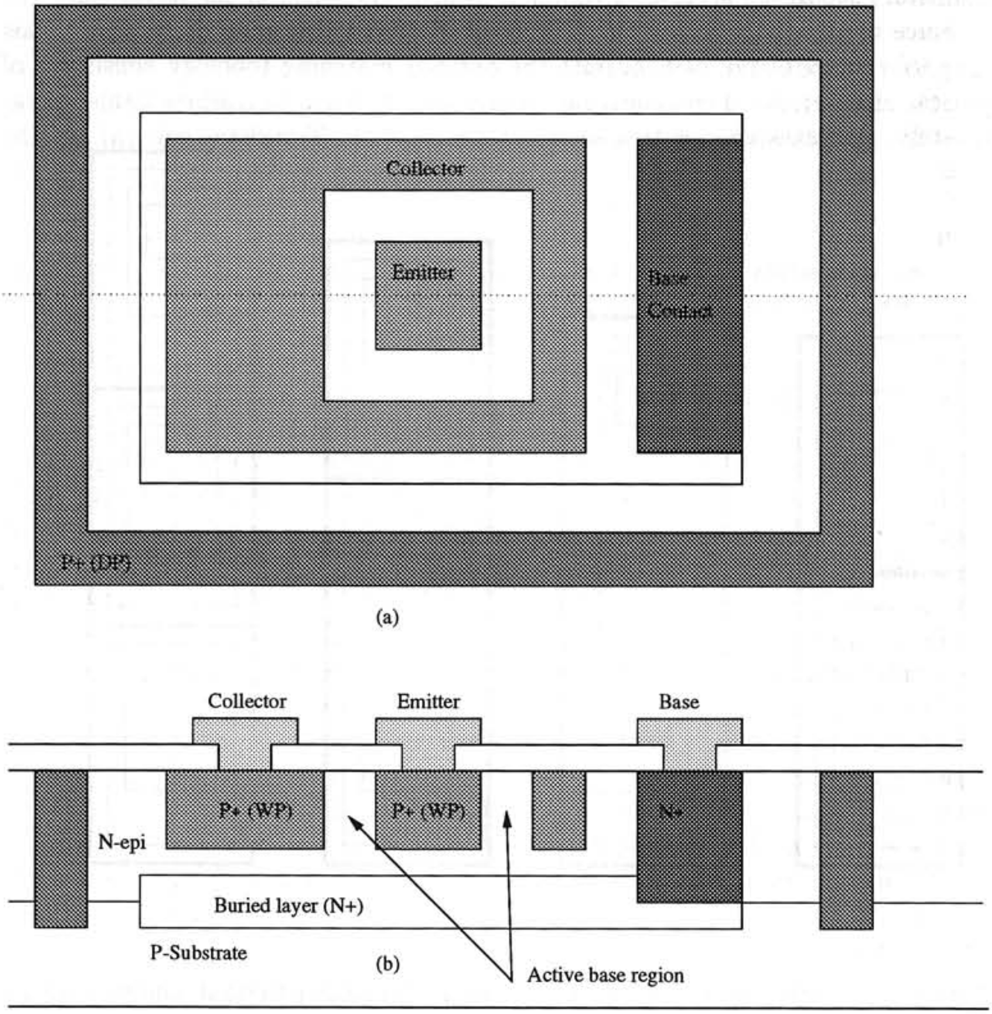


Figure 3.16: A lateral PNP, (a) from the top, (b) a cross-section

The highly doped P layer (WP) is used to implement both the emitter and the collector. The collector encloses the emitter to collect as many of the injected holes as possible. A buried layer is present, to degrade the vertical PNP and to reduce the base resistance of the device.

Note that the capacitor from the substrate to the N region, which was the collector-substrate capacitor for the NPN, is here connected to the base. *The collector has no capacitance to the substrate.* For correct modeling of the influence of this capacitor, the standard SPICE model cannot be used. A separate diode should be used to model the capacitor, or, if available, a special model for lateral transistors should be used. In modern versions of SPICE this model is available.

Since the lateral transistor in the case of the NPN is a parasite that degrades the performance of the average NPN, technology is today focused on reducing the vertical dimensions. This will make lateral transistors finally disappear. Unfortunately, this also implies that the desired lateral PNP is disappearing. Since its performance relies on the perimeter and not on the surface, huge transistors made of comb-like base collector structures are necessary to obtain, for example, a current handling capacity that matches their NPN counterpart. By no means are these transistors complementary to the NPNs. As technology approaches, the need for true vertical PNPs will increase considerably.

3.5 The MOSFET

MOSFETs are transistors of which the operation is based on the flows of majority carriers. This is different from the bipolar transistor of which the operation is based on the injection of minority carriers into the base by the emitter. The latter effect has little dependence on geometry, and the exponential relation between voltage and current depends directly on physical laws. This makes the behavior of bipolar transistors very predictable, which is the reason for the existence of just a few similar models. The specific layout of a bipolar transistor has only influence on the model parameters, and not much on the validity of the model itself.

For MOSFETs there are many effects that may or may not be of significance, depending on the specific layout (size) of the device. Therefore, for each range of operation and each specific layout, a dedicated model may be necessary. Indeed there are many models for MOSFETs, each having their own range of validity. It is not the intention of this book to deal with all the models that exist today. The operation of the MOSFET is shown in a simple way, to obtain some basic insight into its behavior. Some attention is paid to operation in the weak-inversion mode, which is of importance for designing low-power and low-voltage circuits.

3.5.1 MOSFET's basic operation

In Fig.3.17, a cross-section of a MOSFET is shown. The source and the drain

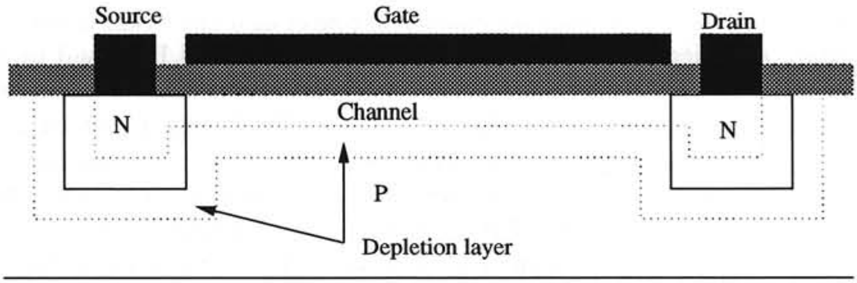


Figure 3.17: Cross-section of a MOSFET, at strong inversion and equal source and drain voltage

voltage are equal ($U_{gs} = U_{gd}$). The source is connected to the substrate (bulk). The substrate is P-type and the drain and source are N-type. For $U_{gs} > 0$, electrons are attracted to the surface. There they form a channel that connects the source and the drain. At a sufficient gate voltage, the electrons become the majority carriers in the area. The channel under the gate has become N-type, which is called *inversion*. The depletion layer between the P and the N region is also shown in Fig.3.17. It bounds the N region, in which the majority carriers are. It can be seen that the depletion layers of the source and the drain have "melted" together. The N region behaves like a resistor. In this mode, the MOSFET could, therefore, also be seen as tunable resistor, tunable via the gate.

With various measures, the voltage at which inversion starts, the *threshold voltage* (VT0), can be set at a desired value. When the threshold voltage is larger than zero, the gate should have some overlap with the source and the drain. Otherwise, in the areas not covered by the gate, a channel could never be generated, and the channel generated under the gate would just be floating without a connection to source or drain. Transistors like this are of the normally-off type. Transistors that do have a channel at zero voltage, the normally-on transistors, need no overlap.

When the drain voltage is raised, the depletion layer around it gets wider, as can be seen in Fig.3.18. This reduces the width of the channel at the drain side. The resistance of the channel is increased. Finally, when at the drain side the threshold voltage is reached, the channel disappears locally. Now, a depletion layer exists between the edge of the channel and the drain. The electric field across this depletion layer is such that an electron that accidentally enters the depletion layer from the channel is driven towards the drain. Depending on the availability of electrons at the edge of the channel, electrons enter the depletion layer at a constant rate, where every available electron is transported to the drain. In this mode, the transistor acts as a current source. The voltage at the drain has no influence on the drain current. The voltage generates the transporting field, but has no influence on the availability of the carriers. It can only transport every

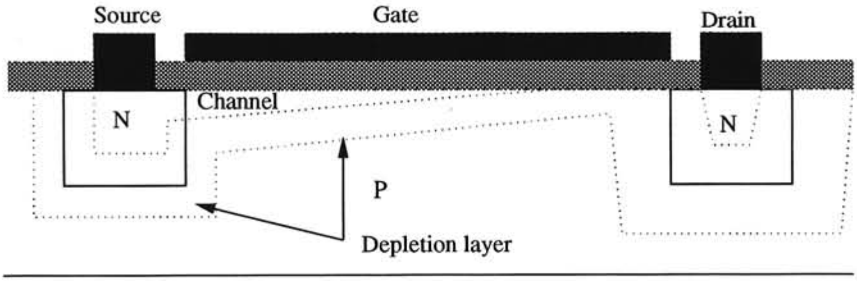


Figure 3.18: Cross-section of a MOSFET, at strong inversion and a raised drain voltage

available carrier, the current is saturated.

As a second-order effect, there is some influence from the drain voltage on the current via the width of the depletion layer between the source and the drain. As the drain voltage is raised, the depletion layer gets wider, so the channel gets shorter. The edge of the channel comes closer to the source, which is an abundant source of electrons, so obviously the availability of electrons at the edge of the channel is greater now. This increases the current.

For P-channel MOSFETs, the behavior is similar and, therefore, it is not discussed in this chapter.

3.5.2 The DC current in a MOSFET

In this section a simple derivation of the current through a MOSFET is given [2]. Though simple, it is correct enough to obtain a sufficient insight into the behavior of the MOSFET. The drain current is given by:

$$I_d = \frac{\text{Charge in the channel}}{\text{Transit time of the charge}} = \frac{Q}{\tau} \tag{3.58}$$

The transit time, the time the charge needs to cross the channel is given by:

$$\tau = \frac{\text{Distance}}{\text{Velocity}} = \frac{L}{\mathrm{v}} = \frac{L}{\mu E} = \frac{L}{\mu}\,\frac{L}{U_{ds}}, \tag{3.59}$$

in which L is the length of the channel, μ is the mobility of the electrons, v is the velocity of the electrons and E is the electric field across the channel.
The charge Q in the channel is found via the capacitance between the channel and the gate and the voltage across it. The capacitance is:

$$C_{channel} = \frac{\epsilon W L}{D} = C_{ox} W L, \tag{3.60}$$

in which W is the width of the channel, ϵ the dielectric constant of the oxide between the gate and the channel and D the thickness of the oxide. The parameters of the oxide are frequently taken together in the parameter C_{ox}.
The voltage across the capacitor $C_{channel}$ depends on the place in the channel. To find the charge on $C_{channel}$, the average voltage is taken:

$$U_{avg} = (U_{gs} - \mathrm{VT0}) - \frac{1}{2}U_{ds} \tag{3.61}$$

From this it follows for the charge:

$$Q = C_{ox}WL\left[(U_{gs} - \mathrm{VT0}) - \frac{1}{2}U_{ds}\right] \tag{3.62}$$

Substituting the results in (3.58) results in:

$$I_d = \mu C_{ox}\frac{W}{L}U_{ds}(U_{gs} - \mathrm{VT0} - \frac{1}{2}U_{ds}) \tag{3.63}$$

In SPICE, μC_{ox} are taken together in parameter KP.

For very small drain voltages the term $\frac{1}{2}U_{ds}$ in (3.63) can be neglected. The equation then yields to:

$$I_d = \mu C_{ox}\frac{W}{L}U_{ds}(U_{gs} - \mathrm{VT0}) \tag{3.64}$$

The drain current now linearly depends on U_{ds}. The MOSFET behaves like a resistor, of which the value can be set via U_{gs}. For larger drain voltages, the term $\frac{1}{2}U_{ds}$ cannot be neglected, it causes (second-order) distortion. This is the reason why many circuits using tunable MOS resistors are implemented as differential circuits. Balancing is a very good measure to counteract even-order distortion.

Equation (3.63) is only valid in the linear region. When pinch-off occurs, thus when at the drain side the channel starts to disappear, the drain voltage loses its influence on the drain current. Pinch-off starts for:

$$U_{ds} = U_{gs} - \mathrm{VT0} \tag{3.65}$$

Substituting this into (3.63) yields to:

$$I_d = \frac{\mu C_{ox}}{2}\,\frac{W}{L}(U_{gs} - \mathrm{VT0})^2 \tag{3.66}$$

To account for the effect of channel length modulation, an extra factor is introduced that models the residual influence of the drain voltage on the drain current. The complete expression for the drain current in the saturation region then becomes:

$$I_d = \frac{\mu C_{ox}}{2}\,\frac{W}{L}(U_{gs} - \mathrm{VT0})^2(1 + \lambda U_{ds}) \tag{3.67}$$

Factor λ is by approximation inversely proportional to the channel length.

3.5.3 Extra effects

There are many models for various simulators. The simplest one is based on the modeling described above. It is the "level 1" model of SPICE. In this model, additionally, the bulk effect, to be described below, is taken into account.

The bulk effect

A MOSFET is actually a four-terminal device, of which the bulk is the fourth connection. The bulk contact is also referred to as the back gate, since it has a similar effect on the channel as the (top)gate. However, this back gate is not isolated from the channel via an isolator, but via a depletion layer. At the back, the MOSFET is actually a junction FET.

The influence of the back gate voltage on the drain current is modeled via the threshold voltage. The back gate voltage U_{bs} influences the threshold voltage via:

$$V_{to} = \mathrm{VT0} + \gamma\left[\sqrt{\phi - U_{bs}} - \sqrt{\phi}\right] \tag{3.68}$$

γ is the named the "bulk threshold parameter" and ϕ the "surface potential". Though they both can be interpreted physically, it is best to consider them as fitted parameters, found by measurement on actual devices and consequent extraction.

The value of γ is given by:

$$\gamma = \frac{\sqrt{2q\epsilon N_B}}{C_{ox}} \tag{3.69}$$

In which N_B is the doping concentration of the bulk. From this it can be seen that the magnitude of the bulk effect depends on the doping of the bulk and can be reduced by doping the bulk lightly. In the extreme case, an isolating substrate results in an absence of the bulk effect.

The bulk effect causes distortion in MOSFETs. It is one of the main causes of distortion in analog MOS circuits. Reducing the bulk doping may reduce the distortion. However, reducing the bulk doping also increases the resistivity of the bulk. This implies that the "ground plane" function of the bulk degrades. Currents flowing through the bulk generate local variations in the bulk potential. These local variation are usually unpredictable and cause, therefore, via the bulk effect, an amount of unpredictability in the threshold voltages of the transistors. This degrades matching, and can even cause undesired coupling between devices.

Velocity saturation

In deriving the expression for the drain current it has been assumed that the velocity of the electrons is unbounded. However, there is a certain critical speed, above which the electrons are not accelerated further. Due to collisions with the lattice, the electrons repeatedly have their drift velocity reset to zero. Between collisions there is no time to accelerate beyond this critical speed. Thus, the relation for the transit time τ given by:

$$\tau = \frac{L}{\mathrm{v}} = \frac{L}{\mu E}, \tag{3.70}$$

The electric field at which velocity saturation occurs has a magnitude of about $3\frac{\mathrm{V}}{\mu\mathrm{m}}$. At higher electric fields the velocity remains limited to v_{max}, a parameter that can be specified in the level 3 SPICE model.

Mobility degradation

The gate voltage gives rise to an electric field that is perpendicular to the movement of the electrons in the gate. This field tends to hamper the movement of the electrons. It reduces their mobility. An increase of the gate voltage and thereby of the perpendicular field introduces a decreasing factor to the drain current. The drain current increases less than it would in the ideal case. Mobility degradation is modeled via:

$$\mu_{eff} = \frac{\mu_0}{1 + \Theta(U_{gs} - \mathrm{VT0})} \tag{3.71}$$

The effect is the other main cause of distortion.

Some remarks on distortion

Two main causes for distortion have been discussed in the previous sections. They are the bulk effect and mobility reduction. The first effect introduces an increasing factor to the drain current, the latter a decreasing factor. It is obvious that both effects could compensate for each other. Measurements have shown, indeed, when there is correct biasing, the overall distortion can be reduced much, and at specific bias point can even be reduced to zero. However, to have zero distortion in a suitable range of operation, specific doping profiles are necessary, and much research will still have to be done to obtain practical results in this way. Still, it is important to remember that there are *two* causes for distortion that *compensate* each other. Removal of one of the effects, for example, removal of the bulk effect by implementing a circuit on an insulating substrate, lets the other cause for distortion work at full strength. It is, therefore, always good practice to have a complete inventory of effects that cause distortion, to prevent anti-distortion measures from "backfiring".

3.5.4 Weak inversion

When the gate voltage is below the threshold voltage, there is no channel. Still some electrons enter the area underneath the gate. This, extremely small, flow of electrons is an injection current, as found in bipolar transistors. The relation between the number of injected electrons and the voltage across the junction is, therefore, also exponential. Once in the region underneath the gate, the electrons diffuse towards the drain junction. This junction, put in reverse bias with an appropriate drain voltage, has a field that brings every electron arriving at the junction to the drain area. This removal of electrons from the drain area generates a concentration gradient, that is the cause of the diffusion current from source to drain underneath the gate. The field across the depletion layer at the drain transports every available electron that appears at its border towards the drain. The magnitude of the current depends on the availability of electrons. This availability is not controlled by the drain voltage, so it can have no influence on the drain current. The transistor behaves like a current source.

For a drain voltage that is sufficiently high, and a source connected to the bulk, the expression for the drain current is:

$$I_d = \frac{W}{L} I_{d0} e^{\frac{U_{gs}}{nkT/q}} \tag{3.72}$$

With $n \approx 1.5$.
This expression closely resembles that of a bipolar transistor. It is though only valid for very small currents.

Again, as a secondary effect, the drain voltage has some influence on the drain current, because it modulates the width of the depletion layer on the drain side. An increasing drain voltage reduces thus the distance electrons have to diffuse from the injection point to the depletion layer at the drain side. This enlarges the concentration gradient of the electrons underneath the gate, and thereby the diffusion current. The magnitude effect is approximately inversely proportional to the length (L) of the gate. This implies that the output impedance of a weak-inversion transistor can be enlarged by increasing its length. Theoretically, there is no limit to the magnitude of the output impedance. In practice, to obtain a useful amount of drain current, the transistor has to be made very wide also, which results in transistors which are gigantic in comparison with the current they deliver.

Standard bipolar transistors, which have comparable properties, cannot be increased in length, because they are vertical devices. Therefore, their output impedance is fundamentally limited. The voltage gain of a weak inversion MOSFET and a bipolar transistor are respectively:

$$A_{\mathrm{MOSFET}} = g_m r_d = \frac{U_E}{nkT/q} L \tag{3.73}$$

$$A_{\mathrm{BJT}} \;=\; g_m r_0 = \frac{\mathrm{VAR}}{kT/q} \tag{3.74}$$

$$\tag{3.75}$$

in which U_E is the Early voltage per unit length of the channel ($U_E \approx 5\mathrm{V}/\mu\mathrm{m}$). Comparing the performance of the two devices, this gives an advantage to weak-inversion MOSFETs.

A problem that exists today, but which will surely be solved in the future, it the stability and accuracy of the threshold voltage. By now, the quality of processing has been improved far enough to make weak-inversion circuits feasible.

3.5.5 Matching

As far as matching is concerned, apart from general rules also found for bipolar transistors, some special measures can be taken with respect to the threshold voltage.
The relation between the difference ΔI_d in the drain current of two equally sized transistors at equal U_{gs} and a difference ΔV_{t0} in threshold voltage is given by:

$$\frac{\Delta I_d}{I_d} = \frac{\Delta\beta}{\beta} \pm \frac{g_m}{I_d}\Delta V_{t0} \tag{3.76}$$

in which $\beta = \mathrm{KP}\frac{W}{L}$.
Since g_m/I_d is inversely proportional to $\sqrt{I_d}$, it follows that the influence of a ΔV_{t0} can be reduced by increasing the drain current. Thus, for example, a current mirror becomes more accurate when it is biased at a large current. Apparently in IC processes with a poorly defined threshold voltage, the currents have to be high to obtain any accuracy. Of course, this strategy is not applicable to weak inversion circuits.

However, from the expression:

$$\Delta V_{t0} = \frac{\Delta\beta}{\beta}\frac{I_d}{g_m} - \frac{\Delta I_d}{g_m}, \tag{3.77}$$

it can be seen that at very low currents, i.e. at weak inversion, a differential pair becomes the most accurate.

3.5.6 The transit frequency

The transit frequency f_T, is defined and also measured in the same way as it is for bipolar transistors. The frequency at which the current gain of a MOSFET equals unity is given by:

$$f_T = \frac{1}{2\pi}\frac{g_m}{C_{gs}} \tag{3.78}$$

g_m is proportional to W/L and C_{gs} is proportional to WL. This implies that the transit frequency is proportional to $1/L^2$. In digital circuits, this it the reason for giving all transistors the minimal gate length. This yields the highest transit frequency and thus the highest transistor switching speed. The current range can be set with the transistor width, without affecting the speed. For analog application, the output impedance is also important in a design. This is worst for the minimal-length transistor. Giving up on the f_T, the output impedance can be increased.

In semi-custom chips designed for the implementation of digital circuits, like a sea of gates, generally no transistors other than minimal-length devices are available. This, of course, greatly hampers the design of analog circuits. The only solution is the use of compound transistors made of a number of transistors in series, to obtain long channel properties. A number of these series connected chains are placed in parallel again, to keep the W to L ratio constant. This may be inefficient, but it is the only way to obtain transistors better suitable for analog design on a "digital" chip.

3.6 Resistors

A resistor is a device that defines a linear relation between a current and a voltage. In the ideal case the relation is:

$$U = R.I \tag{3.79}$$

In (3.79) R is a constant. Thus, R does not depend on time, frequency, temperature, and the value of U and I. In practice, R, unfortunately, also depends on all these factors:

$$U = R(t, \omega, T, U, I)I \tag{3.80}$$

In addition, the relation is also subject to statistical variations due to, for example, process variations. Thus, the behavior of a resistor can be predicted with limited accuracy only.

When an electronic circuit is being designed, usually, first the ideal model according to (3.79) is used. In such a case, the fundamental limits to the performance of the circuit can be found. This performance is compared to the specifications. Then it becomes possible to estimate the magnitude of the errors that can be allowed within the demands of the specifications. From this, the design constraints for the devices can be derived when (3.80) is used.

3.6.1 The value of a resistor

A resistor consists of a "bar" of material with the resistivity ρ and connections to both sides (Fig.3.19). The value of the resistor is determined by ρ, the thickness

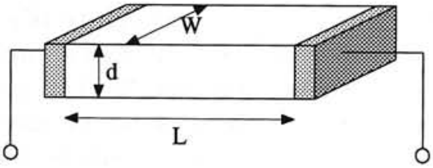


Figure 3.19: A resistor

(d), the length (L) and the width (W).

$$R = \rho \frac{L}{d \cdot W} \tag{3.81}$$

Because, in practice, d cannot be influenced by the designer, only L and W can, ρ and d are usually taken together in a parameter called the *sheet resistance*.

$$R_{\square} = \frac{\rho}{d} \tag{3.82}$$

With the use of this parameter, the layout and value of a resistor is commonly expressed in squares (Fig.3.20). When the number of squares in a resistor is equal

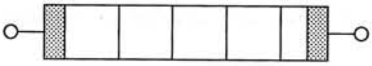


Figure 3.20: A resistor of 4.5 squares

to:

$$N = \frac{L}{W} \tag{3.83}$$

its value can be expressed as:

$$R = N R_{\square} \tag{3.84}$$

From (3.84) it can be seen that it is the number of squares that determines the value of the resistor and not the absolute size of the squares. The maximum current that the resistor can support, its accuracy and its bandwidth, however, do depend on the size of the squares. This is discussed later.

In each technology, there are several layer that are, in principle, suitable for the implementation of a resistor. For each layer, a different number of squares may be necessary to obtain the same resistor value. In table 3.1, for a standard bipolar process, the number of squares necessary to implement a resistor of 1kΩ is indicated.

Suppose the resistor is implemented with the lowly doped P-type layer BW. In Fig.3.21, a cross-section of this resistor is shown. The BW layer cannot be directly connected from the outside. In order to obtain a good contact between the metal

Name	$R_\square$ ($\Omega/\square$)	N
epi	2300	0.43
BW	600	1.67
WP	25	40
metal	0.044	22727

Table 3.1: The number of squares necessary to implement a 1kΩ resistor

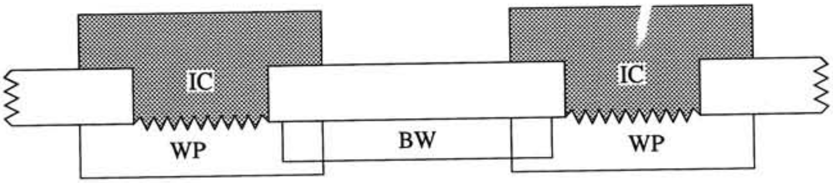


Figure 3.21: Cross-section of a BW-resistor

interconnect layer (IC) and the silicon, a highly doped P layer (WP) is used to interface between IC and BW. Therefore, the resistor actually consists of a series connection of several resistors, as is shown in Fig.3.22. From the left to the right there are:

- The resistance of the metal (IC) from the connection point to the silicon (WP).
  In table 3.1, it can be seen that the sheet resistance of the metal is much lower than that of the BW layer. Therefore, generally, the resistance of the interconnect does not play a significant role.

- The contact resistance between the metal and the silicon.
  The mask that defines the size of the contact opening between IC and WP is the CO mask. The current through the contact opening is a vertical current, as opposed to all other currents in the resistor, which are lateral. Therefore, the resistance of the contact (CO) is not expressed in a sheet resistance, but as a resistance per area. In the standard process used as an example in this book, the resistance of a contact with a size of $2 \times 2\mu$ is about 4Ω. The larger the area is, the smaller the contact resistance will be. When the contact area is enlarged, however, care must be taken that the current flows homogeneously through the contact. If not, the contact resistance becomes larger than expected, and the predictability decreases. When this "current crowding" starts, it is of no use to enlarge the contacts any further.

- The resistance of the WP region.

The sheet resistance of the WP layer is smaller than that of the BW layer, but not small enough to be neglected. Therefore, the number of WP squares in series with the BW resistance should be kept as small as possible.

- The resistance of the BW region.
  This region forms the actual resistor. The number of squares in this layer dominantly determines the resistor vale. Commonly, this part of the resistor occupies most of the area.
- Again the resistance of another BW region, a contact and the interconnect.

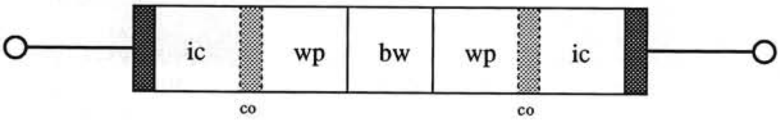


Figure 3.22: The resistor chain of which an integrated resistor consists

From all this, it can be concluded that (3.84) is too simple an expression to design an accurate resistor. This equation yields only the resistance of the BW layer. Therefore, the contribution of the other resistors is added via the contact resistance R_c:

$$R = 2R_c + NR_\square \tag{3.85}$$

It is difficult to calculate the exact value of R_c. However, the order of magnitude in which its contribution lies can easily be estimated. To obtain accuracy, resistors with various BW layer lengths have to be implemented and measured. The value of R_c can be extracted from this data.
The value of R_c can also be determined numerically by a device simulator, but to still supply this simulator with the correct model parameters, accurate measurements have to be done at least once. The results of a simulator are never more accurate than the measurements that were used to generate the model parameters.

Scaling and matching of resistors

Usually, good matching of components is mentioned as one of the big advantages of integrated components. Good matching is, however, not something that happens automatically. Especially when resistors of a different value have to match, have special measures to be taken.
The following rules should be followed insofar as possible.

- The layout for all resistors should be identical. Because R_c matches well within a chip, uncertainty as to its absolute value does not degrade matching.

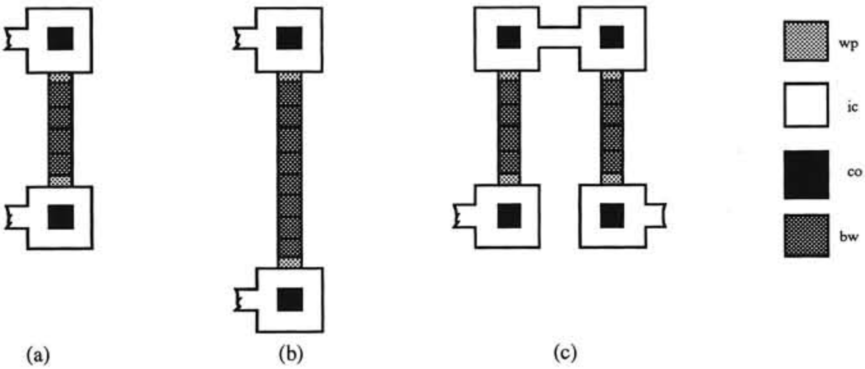


Figure 3.23: Three resistors with a value (a) $2R_c + 4R_{\square BW}$, (b) $2R_c + 8R_{\square BW}$, (c) $4R_c + 8R_{\square BW}$

- Give all resistors the same orientation.
  For example, alignment errors cause the same error in each resistor.

- Place the resistors as close to each other as possible. Variations in sheet resistance across the wafer then have the least influence.

- Pay attention to temperature variations over the chip. When the chip is locally heated, for example by a large dissipating transistor, the resistors should be placed on lines of equal temperature.

- A diffused resistor lies in an epi well, isolated from it by a depletion layer. The width of this layer depends on the junction voltage. Variations in the width of the junctions also cause a variation in the effective width of the layers that make up the resistor and hence of the sheet resistance. This implies that for good matching, all resistors should be insofar as possible at the same potential with respect to the epi layer. When two resistors operate at significantly different DC levels, the voltage on their respective epi wells should be adapted accordingly.

When scaling resistors, care should be taken that the scaling is indeed done linearly. When, for example, two resistors have to be implemented with a value R and $2R$, respectively, for optimal matching, simply doubling the number of squares of the BW layer of the largest resistor does not suffice. From (3.85), it follows that also the contribution of R_c should be doubled. In Fig.3.23 three layouts are shown. It can be seen that the resistor in Fig.3.23c has, accurately, the double value of the resistor in Fig.3.23a, and that for the resistor in Fig.3.23b an error has been made equal to $2R_c$.

Meandered resistors

In the previous sections, it was assumed that all resistors were implemented as straight devices. When the number of squares becomes very large, however, there may not be enough room for a straight resistor to fit on the chip. Also there may be no freedom of choice for the placement of the contacts. In such a case the resistor is meandered. In Fig.3.24, two resistors are shown each consisting of 5 BW squares. Resistor *b* contains one corner. The value of resistor (a) is found with (3.85). Because of the corner, resistor (b) has a different value. A correction has to be made for the corner. In Fig.3.25, it can be seen how the current density varies across the resistor body in the corner. The fact that the current density is not homogeneous across the body, as it is in the straight parts, implies that not all parts of the BW region in the corner equally contribute to the overall resistance. Its contribution is not equal to that of one (straight) square. In Fig.3.25, the local current density is indicated by a vector, of which the *length* is an indication of the magnitude of the current. It can be clearly seen that most of the current flows on the inside of the corner.

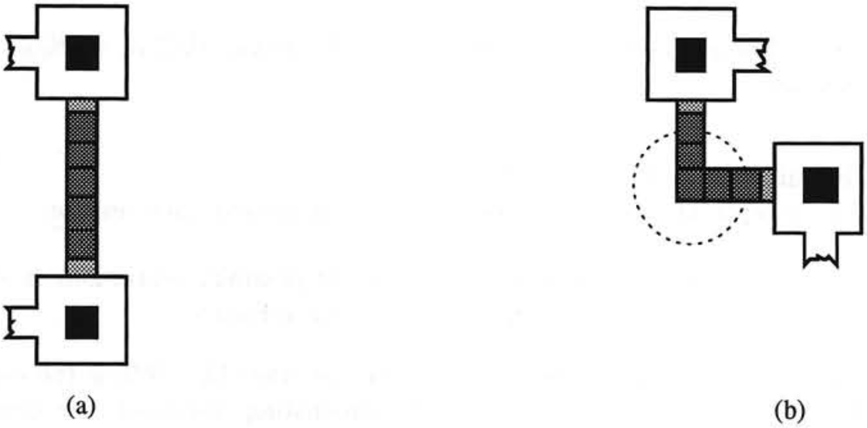


Figure 3.24: A straight and a meandered resistor

Because most of the current takes the shortest path around the corner, the resistance it encounters along this path is also smaller than that of a full-length square. It can be expected that the contribution of a square in a corner is less than that of a square in the straight part of the resistor. Because of this, a correction factor N_k has to be introduces into (3.85). (It will be shown that N_k not only depends on the number of corners, but also on the distance between two corners.) The expression for the resistance then becomes:

$$R = 2R_c + (N + N_k)R_{\Box} \qquad (3.86)$$

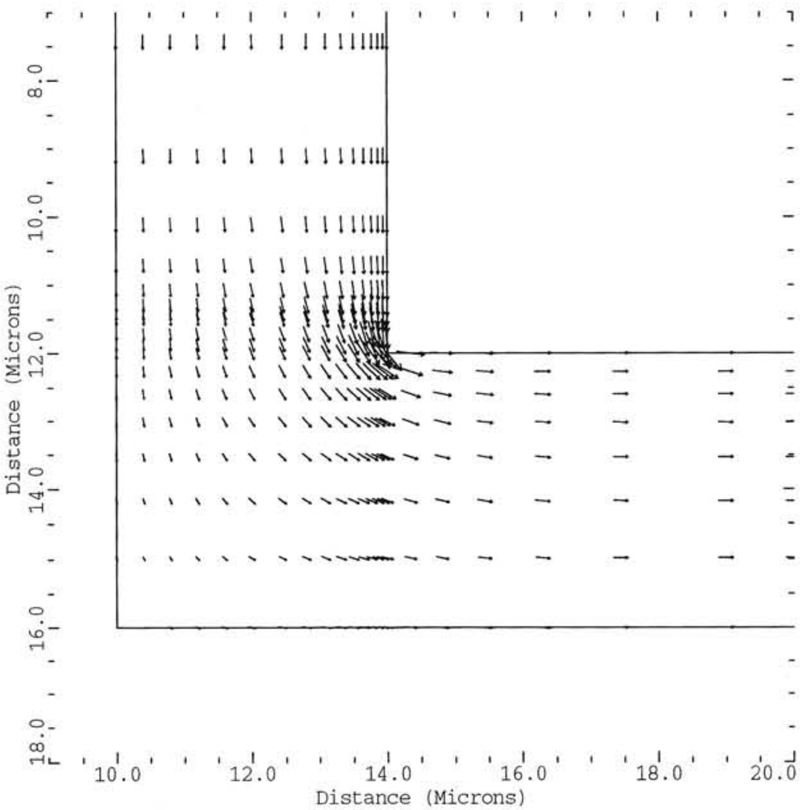


Figure 3.25: The current density in a resistor with a corner (top view). The *length* of the arrows and not their number is a measure of the local current. This figure has been generated with the device simulator PISCES.

in which N is the number of squares in the straight part of the resistor and N_k the effective number of squares contributed by the corners.
For the resistor shown in Fig.3.24b, it follows $N = 4$ and $0 < N_k < 1$.

There are different ways to find the value of N_k. It can be found with the aid of "conformal mapping", a method in which the meandered resistors are transformed into straight resistors by an appropriate transformation of the coordinates. The squares of these straight resistors can then easily be counted. [4]
Generally, these are not the calculations a circuit designer tends to perform. Fortunately, they have to be performed only once for each process. The correction factor can then be found in the design manual.

It is also possible to (mis)use a device simulator like PISCES to find the correction factors. The simulation result given in Fig.3.25 could, for example, be easily used to determine the contribution of a corner. When accurate parameters are available, the simulator results even tend to be more accurate than the theoretical results described above. This is because the actual layout may be difficult to describe mathematically. A corner could be deformed by undesired diffusion as shown in Fig.3.26. In this case, theoretical estimations become very tedious. A device simulator can still produce acceptable results.

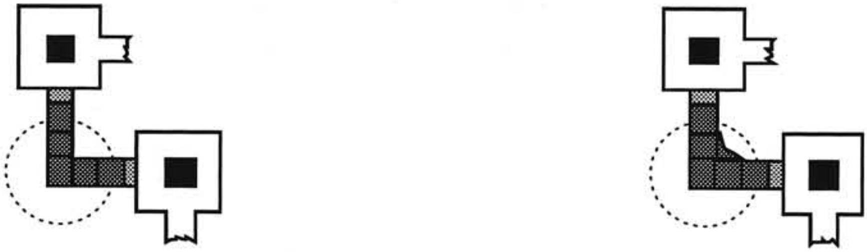


Figure 3.26: Deformation of a corner due to processing artifacts

The third method, and probably the most accurate, is the measurement of dedicated test resistors and the consequent extraction of the parameters.
These measurements are necessary anyway to obtain the model parameters for the methods described above

In table 3.2, for a number of resistors with curves, the contribution of the curves has been given.

Of course, there are many more variations possible than there are shown in table 3.2, but there are hardly any reasons to use others than those of which the data are given in the design manual. A big drawback of "exotic" structures is the lack of accurate data on their behavior. A dedicated characterization procedure has to be started to obtain the data, which costs much time and money.

When accurate matching is required, it is best to avoid corners. Since most current flows through the inside of a corner, it is obvious that the matching becomes

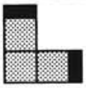
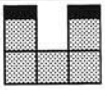
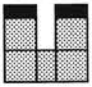
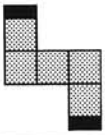
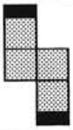
Shape	N	N_k	Value
	2	0.55	$2.55R_\square$
	3	2×0.54	$4.08R_\square$
	2.5	2×0.53	$3.56R_\square$
	3	2×0.54	$4.084R_\square$
	2	2×0.59	$3.18R_\square$

Table 3.2: The correction factors for resistors with curves

very sensitive to any changes of shape (see Fig.3.26) at that place. Therefore, the matching of meandered resistors usually is considerably worse than that of straight resistors.

3.6.2 The structure of a resistor

In Fig.3.27 a cross-section of a resistor in a standard process is shown. It can be seen that the resistor has *four* connections. Two of them are the usual terminals between the desired resistance exists. Between these terminals, the layer that were shown earlier in Fig.3.21 (IC–WP–BW–WP–IC) are found. The contact between the metal (IC) and the silicon (WP) is made via a contact hole. The place and the size of the contact is determined by the (CO) mask.

The resistor is in the epi well. To isolate this well from the resistor, the WP–epi and the BW–epi junctions have to be reverse biased. Because in this case the resistor body is P type, the epi connection should be at a voltage that is higher than the highest voltage that can be expected on the resistor body. The positive supply voltage is a safe choice in this respect. For the connection to the epi layer, a standard NPN-collector contact can be used, since the epi layer forms the collector area of an NPN. The buried layer is also present, because also in this case there is the risk of latch-up via the parasitic PNP.

The epi well is isolated from the rest of the chip via a DP ring around it, and the P substrate at the bottom. To keep the junctions involved reverse biased, the substrate and the ring should be at a sufficiently negative voltage; the negative supply voltage, or ground is the common choice.

There is no need for a separate epi and substrate connection for each resistor. The substrate connection is inevitably a global one, but to have more than one resistor in the epi well is also permissible. They are isolated from each other anyway, because their junctions to the epi layer are reverse biased. This also saves space, because the DP isolation takes much space in comparison to the resistor body itself. Still, for each resistor, the epi voltage should be considered with care.

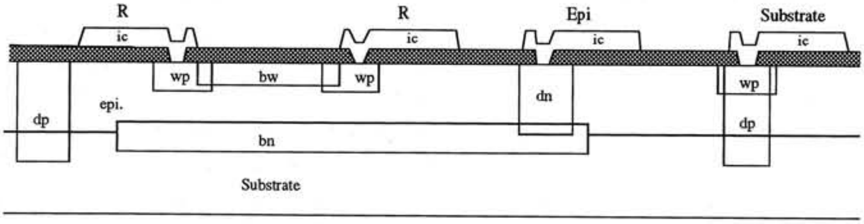


Figure 3.27: Cross-section of a BW resistor

Certainly, in circuits that have more than one supply voltage, it might happen that a resistor is driven at a voltage that is higher than the voltage of its well. Then the junctions between the resistor body and the epi layer become forward biased and the epi layer will "try to follow" the signal voltage at the resistor terminal. Also at that moment, the parasitic PNP starts injecting current into the substrate, since the junction is now biased in forward mode from its base-emitter junction. In Fig.3.28, a model for the resistor from Fig.3.27 is shown. There the connection of the PNP can be seen.

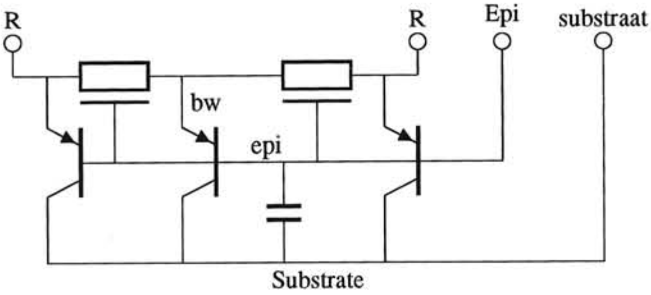


Figure 3.28: A model for a diffused resistor

3.6.3 High value resistors

When high value resistors are to be implemented, they tend to become very long. For example, to obtain a resistor of 100kΩ with the BW layer, according to table 3.1, about 167 squares are necessary. At present, the minimal size of a BW square is about $2\mu \times 2\mu$. This results in a length of about 334μ. To increase the accuracy, the size of the squares may be chosen to be larger, which results in even longer resistors.
Apart from the fact that it is difficult to place resistors like this on a chip, they also suffer from a large parasitic capacitance because of their large area, and therefore tend to have a low bandwidth.

The problems can be solved when layers are made available with a much larger sheet resistance than that of the BW layer. There are a number of ways to obtain such a layer. Sometimes layers are available that have a lower doping level than the BW layer. An example of this is the epi layer. In table 3.1, it can be seen that for the example process, the epi layer has a sheet resistance that is about four times higher than that of the BW layer. Unfortunately, a high sheet resistance alone is not sufficient to obtain small resistors with a high value. In the design manual of the DIMES-01 process, it can be found that for that particular process, the minimal distance between two DP layers, that are used to bound the epi layer is 11μ. This implies that the minimal epi square is $11\mu \times 11\mu$ in size.

This makes the minimal epi square a factor 5.5 larger than the minimal BW square, which more than compensates for the gain in sheet resistance. An epi resistor would become larger than a BW resistor. It can be seen in table 3.3 that the resistor-to-bulk capacitance per square is higher for an epi square than it is for a BW square. For a square with dimensions $W\mu \times W\mu$ it follows:

$$C_{tot} = W^2 C_{bottom} + \mathbf{2} W C_{edge} \tag{3.87}$$

All this shows that a high sheet resistance only is not sufficient to implement resistors of a high value. The parasitic capacitance and the size of the minimal square have to be taken into account too.

Name	C_{bottom} (fF/μ^2)	C_{edge} (fF/μ)	$min.\square$	$C_{min.\square}$ (fF)
epi	0.127	0.42	$11\mu \times 11\mu$	24.61
BW	0.29	0.55	$2\mu \times 2\mu$	3.36
WP	0.31	0.70	$2\mu \times 2\mu$	4.04

Table 3.3: Bottom and edge (junction)capacitance at 0V bias, dimensions of a minimal square and its capacitance

With an extra processing step, it is possible to bring poly-silicon onto the chip which is lowly doped. Then it obtains a high sheet resistance, and, being on top of the silicon rather than in it, the capacitance per square can be rather low. Also, lacking a depletion layer, the voltage dependency is much less than that of the diffused resistors. The price to be paid for this is the extra processing step. Further, there are indications that when the doping is very low, the $1/f$-noise of the resistors is much increased, because of the granular structure of poly-silicon.

The pinch resistor

If the BW layer is suitable to implement resistors, attempts could be made to increase the sheet resistance of the layer locally. Of course, this should be done without changing anything in the processing.

The BW layer is used to implement the base of the bipolar NPN transistor in the process. This implies that in this layer, the emitter (WN) is normally implanted. The emitter is of the highly doped N type. In Fig.3.29a this structure is sketched. It can be seen that underneath the emitter, the BW layer has become very thin. A part of it has been "removed" by the emitter implant. From (3.82) it can be seen that due to this, the sheet resistance of BW under WN is increased considerably. In the DIMES-01 design manual, a separate value of BW under WN is usually given. For the sample process, it is about 7kΩ. Fig.3.29b shows the top

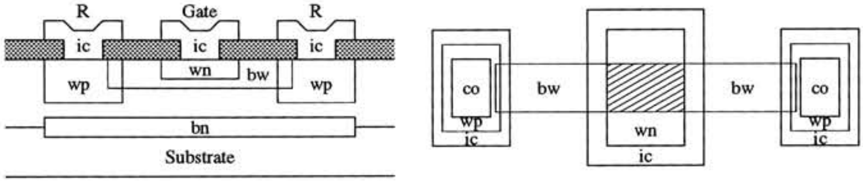


Figure 3.29: Cross-section (a) and top view (b) of a pinch resistor. The shaded part in (b) forms the actual pinch resistor.

view of the resistor. As distinct from the NPN transistor, the emitter (WN) is not enclosed by the BW layer (Fig.3.30b), but is completely intersected (Fig.3.30a). If the intersection were not complete, leakage paths of "thick" BW would be in parallel with the path of "thin" BW. Then, the high resistive part would be short circuited by low-ohmic resistors. The design rules for the NPN transistor are thus neglected on purpose.

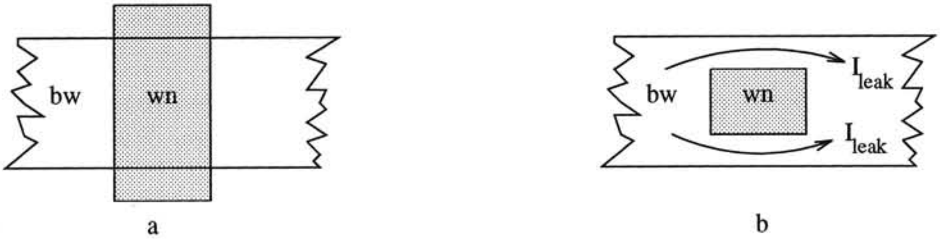


Figure 3.30: (a) WN intersects BW completely (b) WN paled within BW, according to the design rules for the NPN. Low-ohmic leakage paths short-circuit the resistor.

The WN layer is (just like the epi layer) kept at a positive voltage with respect to the resistor body. The junctions WN–BW and epi–BW are both kept in reverse.

The resistor value depends on the reverse voltages across the junctions, because it modulates the width of the depletion layers and thus the size of the resistor body. Actually, the pinch resistor can be seen as a junction FET that operates in the linear region. The only difference is the fact that "real" junction FETs are optimized for optimal transconductance and pinch resistors are not. For junction FETs, the pinch-off voltage is usually lower and better known.
Accurate simulations can, therefore, be done by using the J-FET model in a simulator. In Fig.3.31, it can be seen that the resistor body (channel) is enclosed by depletion layers. The major part of the WN–BW junction extension is, because of the high doping of the WN layer, into the BW region. This causes the resistor to be far more voltage dependent than a standard diffused resistor.

The capacitance per square of a pinch resistor can be calculated with (3.88).

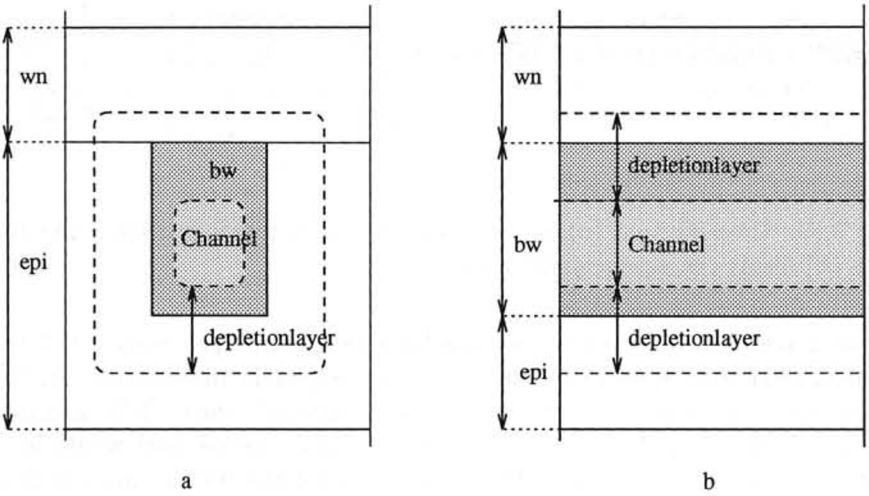


Figure 3.31: Cross-section in the X (a) and Y direction (b) of a section of a pinch resistor. The effective size of the channel is determined by the depletion layers.

The "gate" of the resistor is connected to a constant voltage source, so it can be considered to be grounded. For a square with the dimensions $W\mu \times W\mu$ the capacitance is:

$$C_{tot} = W^2 C_{bottom,epi-BW} + W^2 C_{bottom,BW-WN} + 2WC'_{edge,epi-BW} \qquad (3.88)$$

Note that there is no contribution from a $C_{edge,BW-WN}$. This is because the WN layer is not in the BW layer as is usual, but it intersects the BW layer. Therefore, the value $C_{edge,epi-BW}$ found in the design manual is also an overestimation.
With:

$$\begin{aligned} C_{bottom,epi-BW} &= 0.29fF \\ C_{bottom,BW-WN} &= 1.9fF \\ C'_{edge,epi-BW} < C_{edge,epi-BW} &= 0.55fF \\ W &= 2\mu \end{aligned}$$

all parameters taken from the DIMES-01 design manual, it follows $C_{tot} = 10.96fF$. Thus, when the capacitance of a pinch resistor implemented in the sample process is calculated, it appears it is much less than that of an equivalent BW resistor, because much less squares are necessary. The capacitance per square is about a factor 3 higher than that of a BW square, but the resistance is a factor 12 higher. Therefore, the bandwidth of a pinch resistor can be expected to be higher than that of its normal diffused equivalent. Its size is also considerably smaller and, therefore, pinch resistors are perfectly suitable for low power and low voltage circuits. Their non-linearity is mild, and it can be substantially compensated for by using differential circuit topologies.

References

[1] H.C.de Graaf and F.M.Klaassen, *Compact Transistor Modeling for Circuit Design*, ISBN 3–211-82136–8, Springer-Verlag, Wien–New York, 1990.

[2] Y.P.Tsividis, *Operation and Modeling of the MOS Transistor*, McGraw-Hill Book Company, New York, 1987.

[3] Peter Ashburn, *Design and Realization of Bipolar Transistors*, John Wiley and Sons Ltd., November 1992.

[4] P.M.Hall, *Resistance calculations for thin film patterns,* Thin Solid Films, 1 (1967/68) pp. 277-295, Elsevier, Amsterdam.

[5] R.P.Nanavati, *Semiconductor Devices*, Intext Educational Publishers, 666 Fifth Avenue, New York, 1975.

[6] S.M.Sze, *Physics of Semiconductor Devices*, John Wiley and Sons, New York, 1969.

[7] G.G.E. Low. *Carrier concentration disturbances in semiconductors*, Proceedings of the Physical Society B, 68:310–314, 1955.

[8] R.S. Muller and Th.I. Kamins. *Device electronics for Integrated Circuits*, John Wiley & Sons, New York, 1986.

Chapter 4

Translinear Circuits

Albert van der Woerd

4.1 Introduction

4.1.1 About the term "translinear"

Early translinear circuits were strictly based on the remarkable fact that the *trans*conductance of a bipolar transistor is *linear*ly proportional to its collector current. This fact is a consequence of the logarithmic relation between I_C and V_{BE}

$$V_{BE} = V_T \ln(I_C / I_S(T)) \tag{4.1}$$

from which it follows

$$\frac{\delta I_C}{\delta V_{BE}} = g_m = \frac{I_C}{V_T} \tag{4.2}$$

This is the key to the strictly translinear principle and, basically, only devices showing a very exact logarithmic relation are suitable. As MOSTs operating in weak inversion show a comparable relation between the gate-source voltage and the drain current, they are suitable too for application of the strictly translinear principle.

Distinction between Translinear Loops (TL) and Translinear Networks (TN)

A general property of a TL circuit is that it contains one or more closed loops of emitter-base junctions (gate-source junctions) with a voltage-current relation according to (4.1). A TN circuit, however, contains no such loops but yet the relations (4.1) and(4.2) are intentionally and profitably used. In practice, configurations are often combinations of TL and TN circuits.

However, over time the term "translinear" has come to refer to a wider class of circuits, for some "translinear" circuits contain MOSTs in strong inversion, whose

$V_{GS} - I_D$ relation is quadratic instead of exponential, whereas the other properties (the presence of loops of gate-source junctions and/or the intentional use of the $V_{GS} - I_D$ relations) are maintained.

4.1.2 General classification of translinear circuits within the world of analog circuits

As stated before, all analog circuits where the logarithmic voltage-current relation of individual devices is intentional and is profitably used are called Translinear Networks (TN). However, a special class of circuits, containing one or more closed loops of junctions (TL), needs further consideration.

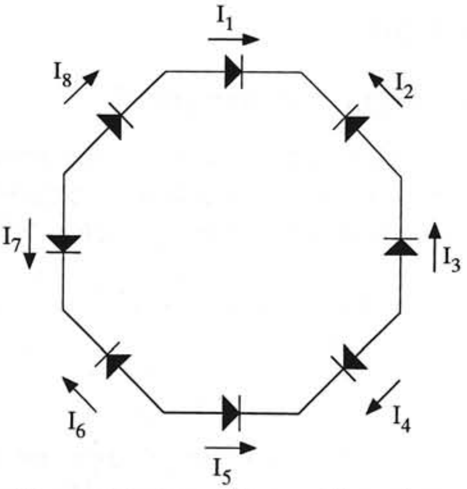


Figure 4.1: Closed loop of junctions

To begin with we consider the closed loop of junctions shown in Fig. 4.1. We assume that all junctions (which can be diodes or the input ports of transistors) are forward-biased with circuitry that is not shown. Other boundary conditions are that the loop must contain an *even number of junctions (at least two)* and that there are an *equal number of junctions clockwise facing and counterclockwise facing* (shorted CW and CCW). If the forward voltage of each junction $(1, 2, 3, ...n)$ in Fig. 4.1 is V_{FK}, it follows

$$\sum_{k=1}^{k=n} V_{FK} = 0 \tag{4.3}$$

If we assume that V_T is device-independent and the collector (drain) current of the device with junction number k is I_k , (4.1) and (4.3) yield

$$\prod_{k=1}^{k=n} \frac{I_k}{I_{sk}} = 1 \tag{4.4}$$

If we distinguish the clockwise and counterclockwise facing junctions (each $n/2$) we can write

$$\prod_{CW} \frac{I_k}{I_{sk}} = \prod_{CCW} \frac{I_k}{I_{sk}} \tag{4.5}$$

Since for bipolar transistors the saturation currents I_{sk} are proportional to the correspondent emitter areas A_k, we can rewrite (4.5) as

$$\prod_{CW} \frac{I_k}{A_k} = \prod_{CCW} \frac{I_k}{A_k} \tag{4.6}$$

or as

$$\prod_{CW} J = \prod_{CCW} J \tag{4.7}$$

where J is the saturation current density of each junction.

Equation (4.7) gives the ultimate translinear principle, in words (Gilbert [1]): *In a closed loop containing an even number of forward biased junctions, arranged so that there are an equal number of clockwise facing and counterclockwise facing polarities, the product of the current densities in the clockwise direction is equal to the product of the current densities in the counterclockwise direction.*

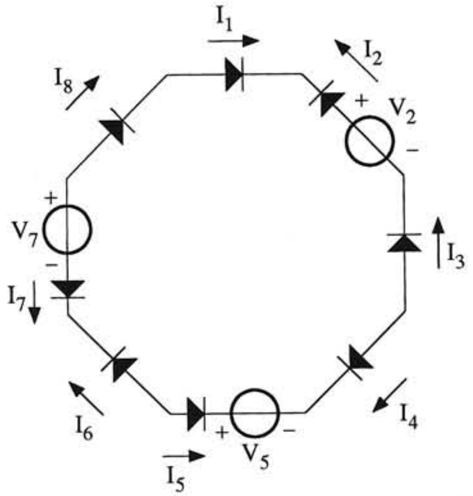


Figure 4.2: Closed loop of junctions and some voltage sources

4.1.3 Extension of TL theory to include dc voltage generators

Fig. 4.2 depicts a modification of the general circuit shown in Fig. 4.1, where some dc voltage sources have been added. If V_L is the net voltage in the loop, the modified form of Eq. (4.3) is

$$\sum_{k=1}^{k=n} V_T \ln \frac{I_k}{I_{sk}} = V_L \tag{4.8}$$

A few practical circuits using TL loops including dc voltage sources will be shown in Section 3.

4.1.4 Application areas of translinear circuits

Linear amplification must be considered to be the most important class of analog signal processing. Though Gilbert introduced in 1968 "a new wide-band amplifier technique" based on the translinear idea, later developments in structured amplifier design have shown that "basic" amplification is surely not the most powerful translinear circuit issue. Because indirect feedback techniques are used in translinear circuits such a design does not have optimal noise, accuracy and linearity qualities [2]. However, if the gain must be *controllable*, the translinear principle can yield great advantages. Further, they have proved to be very powerful for providing a great amount of different *nonlinear* signal processing functions, such as analog multiplying/dividing, rms-dc conversion, vector summation, squaring and square-rooting. A general restriction is that the special features of the translinear principle are only obvious at relatively low frequencies. At high frequencies other principles to provide nonlinear signal processing are often more powerful.

A special class of analog electronics that has attracted much interest during the last few decades is the design of *low-power/low-voltage* circuits. In this class we observe a revival of some types of translinear circuits. This is mainly because the *current-mode* operation of translinear circuits perfectly fits with *low-voltage* operation, whereas *low-power* operation generally implies that the system bandwidth is restricted. (Note, that the low-frequency area is the most powerful operation area of TL circuits).

4.1.5 Suitable semiconductor components

If we only consider circuits operating according to the *strict* translinear principle (Eq. 4.7), we must resort to devices with a perfectly exponential transfer. BJTs fulfill this requirement within a very large collector current range. Fig. 4.3 gives an example of the measured $I_C - V_{BE}$ characteristic of a typical BJT.

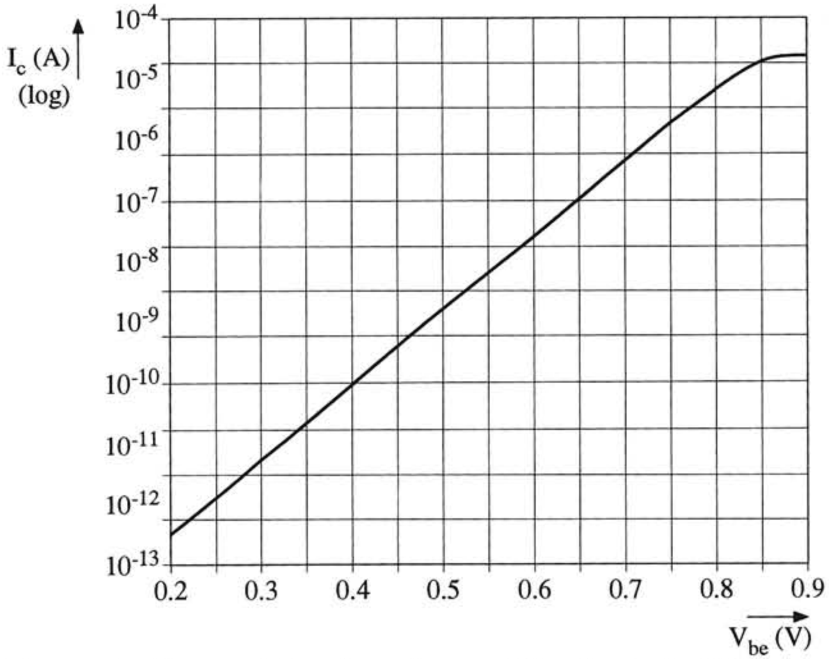


Figure 4.3: Measured $I_C - V_{BE}$ characteristic of a typical BJT

Other suitable devices for the strictly translinear principle are MOSTs operating in weak inversion. Fig. 4.4 shows the $I_D - V_{GS}$ characteristic of a typical MOST.

If the strictly translinear principle is no longer maintained, MOSTs operating in moderate/strong inversion are also suitable.

4.2 Design strategies for translinear circuits

4.2.1 The heuristic approach

The term "heuristics" literally means "method of solving problems by inductive reasoning, by evaluating past experience and moving by trial and error to a solution". The first design approaches of most known electronic circuits were done in this way and, consequently, generally only experienced engineers are able to find new solutions by using this approach.

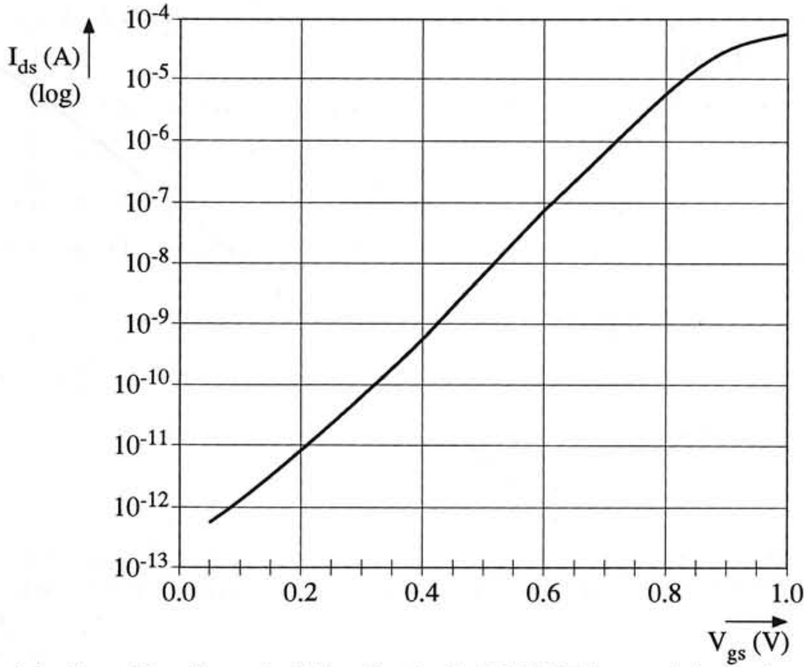


Figure 4.4: $I_D - V_{GS}$ characteristic of a typical MOST in weak inversion

4.2.2 The systematic approach

A systematic design system must contain a set of generally valuable, structured design rules. These rules must be structured in a hierarchical way, so that, from a restricted set of suitable basic configurations, all possible solutions to a preliminary stated problem are generated. The approach has successfully been applied to the design of amplifiers with overall feedback and also to translinear circuits. The advantages of this approach are twofold. First, a well-structured design system can be used by designers without specialized talent or experience. Second, it generally generates more (and sometimes better) solutions to the same problem than would have been found by heuristic designing.

However, systematic design systems have some serious drawbacks too. Generally, the solutions generated by such systems preferably must be selected by an experienced designer, first because not all solutions are practically appreciable and second because some solutions don't work at all. The last phenomenon is because the system generally is not able to process all electrical properties. As an example: in synthesis systems for translinear circuits some resulting circuits may show positive feedback loops (resulting in oscillation or latching) because the system is not able to recognize this item.

4.2.3 Interaction between the heuristic and the systematic approaches

The development of systematic design systems has always been the result or continuation of much work carried out in a heuristic way. They are valuable to generalize and complete the heuristically found solutions. Therefore, the importance of heuristic reasoning should never be depreciated. However, new, systematically found solutions can deliver new impulse and fresh understanding to the heuristically reasoning designer.

4.3 Examples of heuristically found TL configurations

4.3.1 General

This Section deals with the most well-known TL circuits. All of them were initially designed for realization in a bipolar process, and hence some quality parameters are coupled with the influence of finite base currents. To date, the quality standards of these devices are according to the state of the art. However, as soon as comparable circuits are designed with MOSTs in weak inversion, some quality standards will undoubtedly have to be revised. As an example, in Section 4.3.3 the "beta-immune" type-A analog multiplier cell is considered to have better linearity and accuracy than the "beta-sensitive" type B cell. However, with MOSTs operating in weak inversion, which have infinite "betas", this comparison is senseless. As still little is known about the quality aspects of standard TL circuits operating with MOSTs in weak inversion, we resort to bipolar circuits in this Section.

4.3.2 Current mirrors

Note: We only deal with the basic current mirror form here from a translinear viewpoint. Numerous developments of the basic form, with properties suited to special applications, have been made. They are extensively dealt with in Chapter 5 of this book.

Fig. 4.5 shows the simplest current mirror. If the general TL equation (4.6) is applied to this circuit we observe that there is one translinear loop with only one CW junction and one CCW junction, yielding

$$\frac{I_{C1}}{A_1} = \frac{I_{C2}}{A_2} \tag{4.9}$$

If the circuit is fed with an input current I_{in} , disregarding the Early effect, and assuming $h_{FE1} = h_{FE2}$, some calculation with (4.9) and the relation of a BJT:

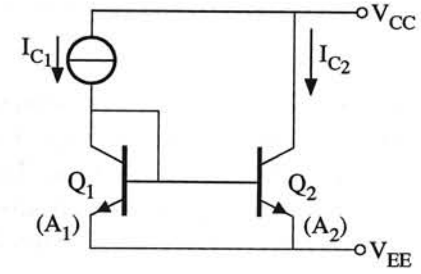


Figure 4.5: Simple current mirror

$I_C = h_{FE} I_B$ yields

$$I_{C2} = \frac{I_{in}}{\frac{A_1}{A_2}\left(1 + \frac{1}{h_{FE}}\right) + \frac{1}{h_{FE}}} \tag{4.10}$$

Hence, the circuit can be considered to be a dc current amplifier or attenuater with a gain that depends on the value of A_1/A_2 and an error that depends on the influence of the finite value of h_{FE}.

If both transistors are biased with appropriate collector currents (Fig. 4.6), class-A current amplification/attenuation is possible too. In that case I_{C2}, I_{in} and h_{FE} in (4.10) must be replaced by i_{out}, i_{in} and β_{ac}, respectively.

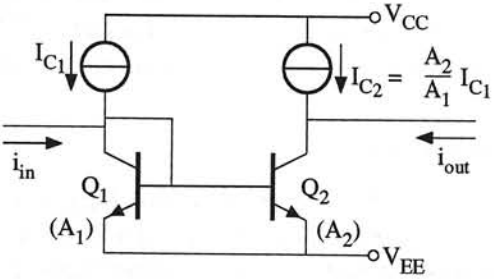


Figure 4.6: Current mirror as a class-A amplifier/attenuater

4.3.3 (Controllable) amplifiers and attenuaters; analog multipliers

Note: Because analog multipliers can be used as controllable amplifiers by replacing one of the input signals by a control signal, they are dealt with together.

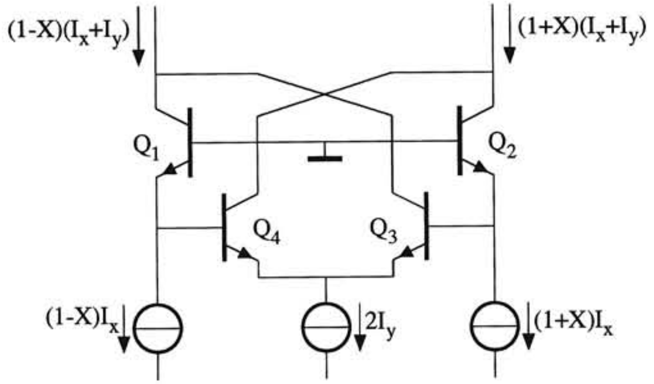


Figure 4.7: The "Gilbert Gain Cell"

The "Gilbert Gain Cell"

Fig. 4.7 depicts a circuit that is suitable for controllable current amplifiers. The transistors Q_1 through Q_4 have equal emitter areas and form a translinear loop. Using Eq. (4.6) we have

$$I_{C1}I_{C4} = I_{C2}I_{C3} \tag{4.11}$$

First the base currents are disregarded. The circuit is symmetrically fed with input currents I_i^+ and I_i^- which are the emitter currents of Q_1 and Q_2. To get suitable expressions for the output currents I_o^+ and I_o^- we introduce a *modulation index*, X, and write

$$I_i^+ = (1+X)I_X \tag{4.12}$$
$$I_i^- = (1-X)I_X \tag{4.13}$$

with $-1 < X < +1$. Combination of (4.11) and (4.13) yields

$$I_o^+ = (1+X)(I_X+I_Y) \tag{4.14}$$
$$I_o^- = (1-X)(I_X+I_Y) \tag{4.15}$$

Example: If I_Y is chosen 9 times I_X, the circuit will show a (differential) current gain of 10.

Note: The above-introduced notation method with one or more modulation indexes will be frequently used in the following.

Errors caused by finite base currents $I_{B,Q3,4}$ is (much) larger than $I_{B,Q1,2}$. These currents are added to the output currents via Q_1 and Q_2, but in anti-phase. Therefore, the error in the output currents due to base currents is considerable and the maximum gain is in practice limited to about 10.

The types A and B translinear cells; two-quadrant multipliers

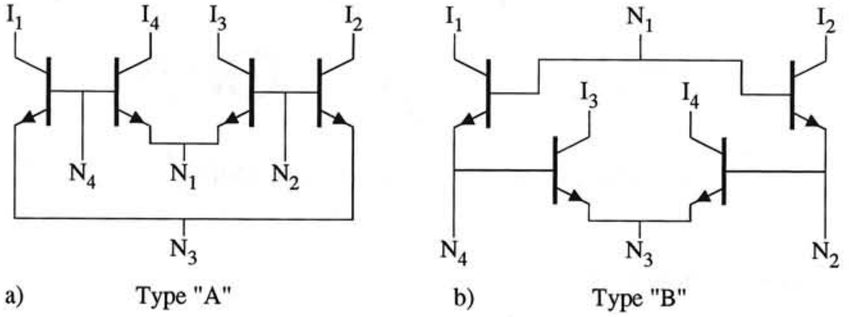


Figure 4.8: Types "A" and "B" TL cells

If four junctions are series-circuited in a closed TL loop there are two possibilities, depicted in Fig 4.8. Left, the junction polarities are alternating and right they are balanced. Therefore, the structures in Fig. 4.8 are referred to as the "A" and "B" cell types. We now take a closer look at both cells.

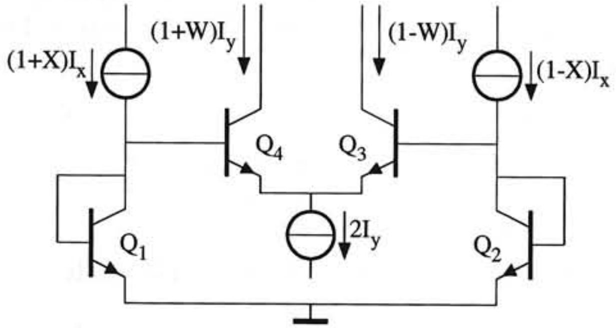


Figure 4.9: The "A" cell with biasing

The "beta-immune" Class-A cell Fig. 4.9 gives the cell, completed with suitable in- and output signals. The translinear condition (Eq. 4.6) yields

$$I_1 I_3 = I_2 I_4 \tag{4.16}$$

or equivalently

$$\frac{I_1}{I_2} = \frac{I_4}{I_3} \tag{4.17}$$

From (4.16) and (4.17) it is immediately apparent that the cell can be used as a one-quadrant analog multiplier or divider by choosing one of the currents to be constant. However, we will show now, that the circuit is also suitable as a two-quadrant multiplier. Therefore, we apply a modulation index to the currents I_1 through I_4, so that

$$(1+W)I_y(1-X)I_x = (1-W)I_y(1+X)I_x \tag{4.18}$$

From (4.18) it is apparent that $W \equiv X$ *for any value of X between -1 and +1, irrespective of β, transistor geometry and temperature!* This is the reason why this cell is called "beta-immune". If the differential output current is called $I_w = (1+W)I_y - (1-W)I_y$, we obtain

$$I_w = 2XI_y \tag{4.19}$$

Hence, the cell can operate as a two-quadrant multiplier. Since the differential input signal is $2XI_x$, the current gain is just I_y/I_x.

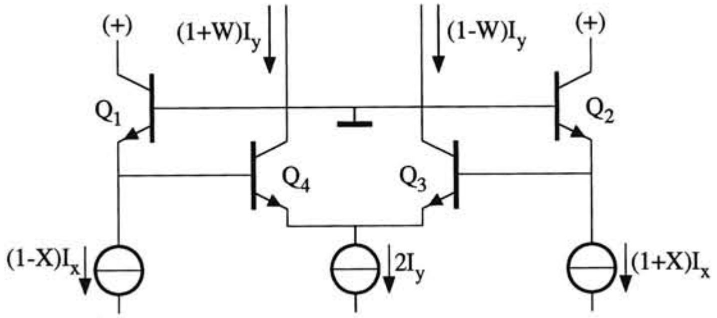


Figure 4.10: The "B" cell with biasing

The "beta-allergic" Class B cell Fig. 4.10 shows the basic cell as a two-quadrant multiplier. It is somewhat more affected by finite beta than the Class A cell. To demonstrate this, we introduce a "base current defect factor" δ to represent either $I_B/I_C(=1/\beta)$ or $I_B/I_E(=1/(\beta+1))$ [3]. Applying the translinear relation (4.7) to Fig. 4.10 now yields

$$(1+W)I_y((1-X)I_x + \delta(1+W)I_y) = (1-W)I_y((1+X)I_x + \delta(1-W)I_y) \tag{4.20}$$

or

$$W = \frac{X}{1 + \delta \frac{I_y}{I_x}} \tag{4.21}$$

Hence, now the modulation index W is not an exact replica of X, thus resulting in some errors caused by finite betas.

Four-quadrant multipliers

The Class-B cell as a four-quadrant analog multiplier With suitable input signals and summing the collector currents of $Q_{1,4}$ and $Q_{2,3}$ respectively, the Class-B cell can also be used as a four-quadrant multiplier. The circuit with in- and output currents is shown in Fig. 4.11. Applying the translinear relation (4.7) we find

$$I_4 I_1 = I_2 I_3 \tag{4.22}$$

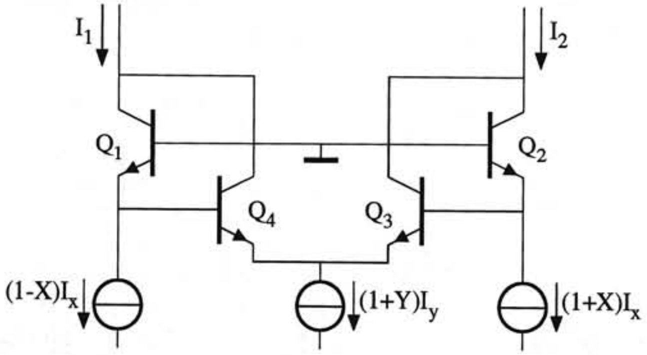


Figure 4.11: The "B" cell as a four-quadrant multiplier

Further it appears from Fig. 4.11

$$I_3 + I_4 = (1 + Y) I_Y \tag{4.23}$$

From (4.22) and (4.23) we easily find

$$I_3 = \frac{(1+Y)(1-X)I_Y}{2} \tag{4.24}$$

and

$$I_4 = \frac{(1+Y)(1+X)I_Y}{2} \tag{4.25}$$

which results in

$$I_{out} = I_5 - I_6 = -2XI_X + XYI_Y + XI_Y \tag{4.26}$$

If I_Y is chosen $2I_X$ the final result is

$$Z = \frac{i_{out}}{I_Y} = XY \text{ for } -1 \leq X, Y \leq +1 \text{ (four-quadrant operation)} \tag{4.27}$$

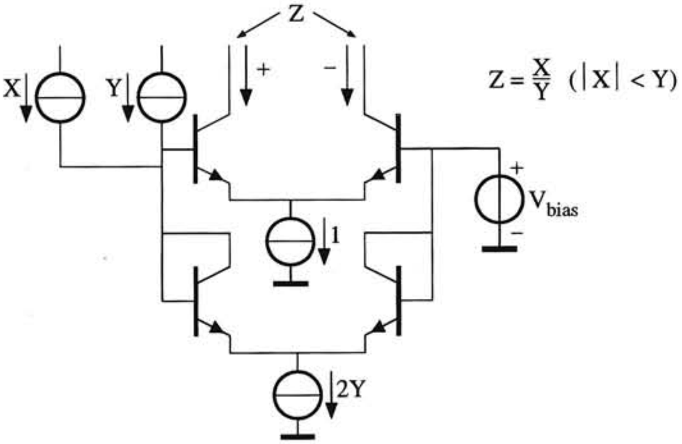


Figure 4.12: Class-"A" cell as an analog divider

Operation as a squarer, as an analog divider or as a square-rooter If X and Y are fed with the same input signal, the cell acts as a squarer. Further, two-quadrant dividing is realized if the functions of one of the input signals (X or Y) and the output signal Z are reversed. Fig. 4.12 shows a possible implementation of a class A cell, where the variables are simply identified with the (normalized) currents. (This circuit was found by employing the systematic design method, outlined in 4.4.2). However, since $|X| < Y$, its dynamic range is restricted. If a translinear multiplier is placed in the feedback loop of a current amplifier with two identical outputs, the dynamic range can be enlarged (Fig. 4.13). Finally, square-rooting is provided if $X = Y$ in Fig. 4.13.

The "six-pack" translinear four-quadrant multiplier The best-known four-quadrant translinear core (dubbed the "six-pack") is shown in Fig. 4.14. This circuit is completely balanced and contains two overlapping Class-A cells. The reader is invited (in the same way as in Fig. 4.11) to find the expression for $I_{out} = I_7 - I_8$ as a function of X, Y and the biasing currents I_X and I_Y.

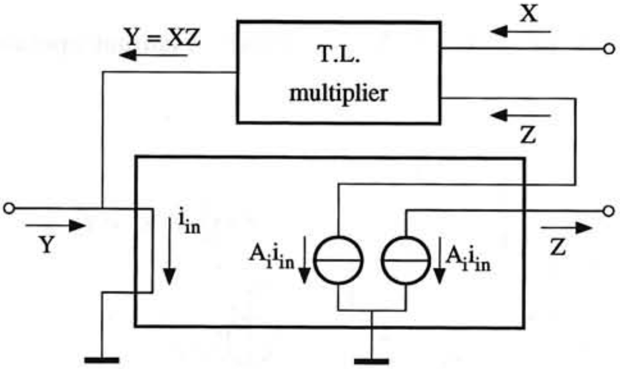


Figure 4.13: High-range analog divider

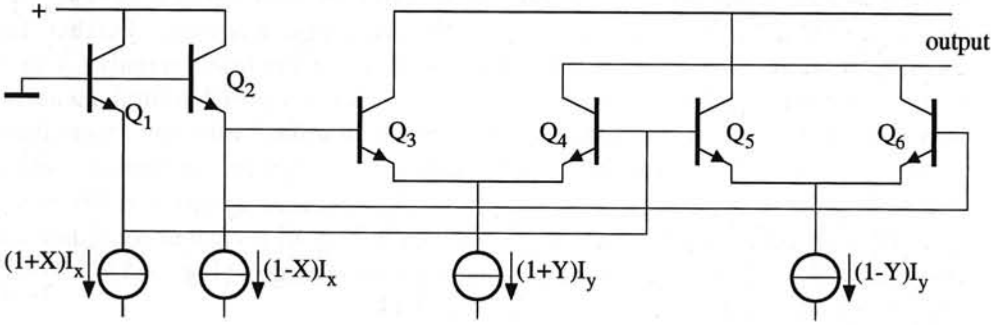


Figure 4.14: The "six-pack" four-quadrant multiplier

4.3.4 The translinear "cross-quad" and its applications

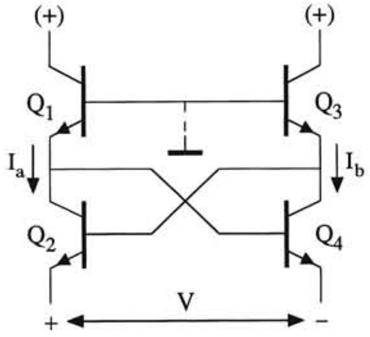


Figure 4.15: The basic TL cross-quad

The basic core Though it is not strictly TL (the loop is broken, making it a TN form) it is closely related and has many uses, both by itself and embedded in TL circuits. Fig. 4.15 depicts the generic cell. Say that, in some way, a current I_a is established in $Q_{1,2}$ and I_b in $Q_{3,4}$. The voltage in the open port is

$$V = V_{BE4} + V_{BE1} - V_{BE3} - V_{BE2} \tag{4.28}$$

Hence, if the base currents are disregarded, V is always zero, irrespective of the values of I_a and I_b. In a real circuit V can come quite close to zero. It's easily shown that for $\beta = 100$, I_a is fixed and I_b is swept from $0.1I_a$ to $10I_a$, $V_{\max}$ would be $\pm$ 2.5 mV. Conversely, $V_{\max}$ could be viewed as the voltage required to establish a 10:1 current ratio in the two transistor pairs. Note that this is much lower than the "60mV per decade", associated with a simple pair of junctions. The input resistance with $I_a = I_b = I$ is found to be

$$R_{in} \approx \frac{4V_T}{\beta I} \tag{4.29}$$

One drawback of the cell is the fact that the radical reduction of the input resistance is obtained by 100% positive feedback, and the circuit is prone to oscillation if not correctly used. Nevertheless, the circuit has many useful applications. Some of them will be included here.

Caprio's Quad

By applying a voltage source between the bases of Q_1 and Q_3 as shown in Fig. 4.16 (dubbed Caprio's quad [3]), this voltage will be replicated across the resistor R. Hence, owing to the V_BE cancellation, the circuit operates as an accurate and

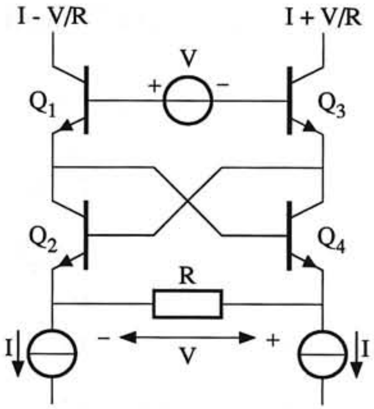


Figure 4.16: Caprio's quad

linear transconductance. A major drawback is that the circuit has a negative input resistance ($(-R)$) between the bases of Q_1 and Q_3 (the reader is invited to prove this). This can easily lead to instabilities if the source is slightly reactive. Fig. 4.17 shows a special application: a linear half-wave rectifier results, if one of the current sources (I) is chosen to be zero. Fig. 4.18 depicts the simulated transfer $I_C(Q2) = f(V)$.

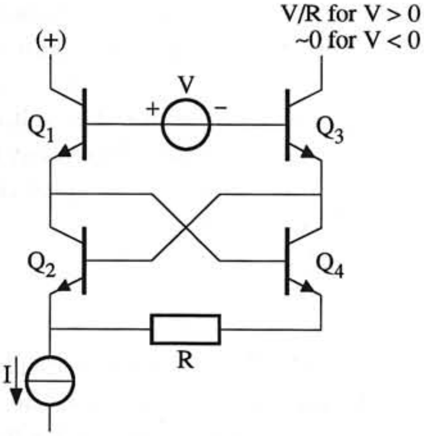


Figure 4.17: Half-wave rectifier derived from Caprio's quad

The translinear cross-quad as a PTAT cell

Most well-known PTAT generators are self-supporting: they produce a PTAT current without external biasing current. Because those cells show a second sta-

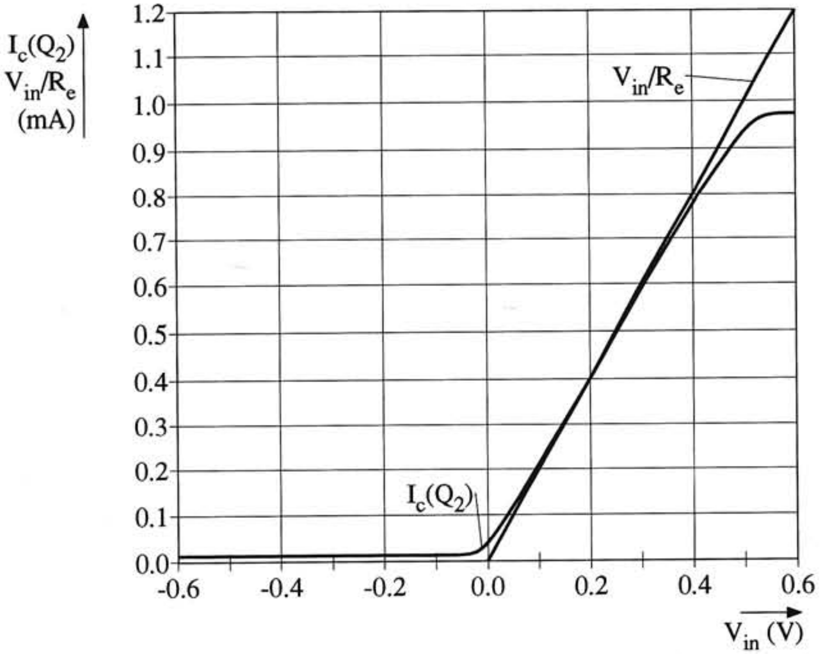


Figure 4.18: Simulated transfer of the half-wave rectifier

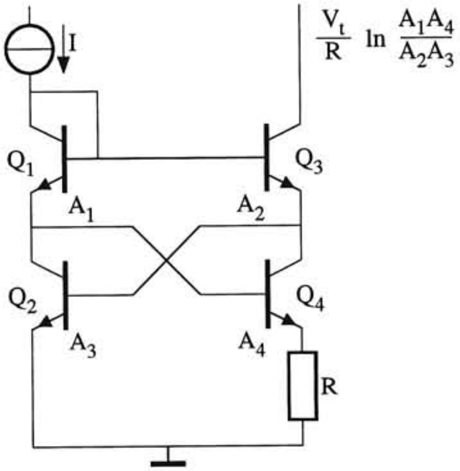


Figure 4.19: PTAT generator with TL cross-quad

ble biasing position, where all currents are zero, a starting circuit is generally indispensable. Fig. 4.19 depicts a PTAT cell with a translinear cross-quad. Consideration of the TN path shows that

$$I_{C3} = \frac{V_T}{R} \ln \frac{A_1 A_4}{A_2 A_3} \tag{4.30}$$

Hence, the PTAT current is independent of the input current I, which can be considered as a (non-critical) starting current.

Operation in class-B (AB); applications in power amplifiers

From the translinear relation (4.6) it appears that in each translinear loop the small signal operation is independent of the loop current. Hence, all current relations are maintained, if the input and output signals are directly used as biasing currents, provided that they remain > 0 (class-B operation). This means that all true translinear circuits are basically suitable for class-B operation. The most well-known application is the traditional complementary class-AB power amplifier, shown in Fig. 4.20.

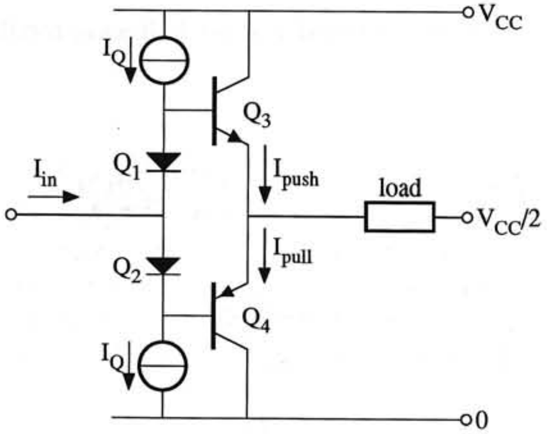


Figure 4.20: Basic class-AB output stage

$Q_{3,4}$ are the output (power) transistors and $Q_{1,2}$ are the drivers. The addition of a quiescent current I_Q (dashed), which is always smaller than the peak values of the input current, makes it class-AB operated.

Applying the translinear law in Q_1 through Q_4 yields

$$I_{\text{push}} I_{\text{pull}} = I_Q^2 \tag{4.31}$$

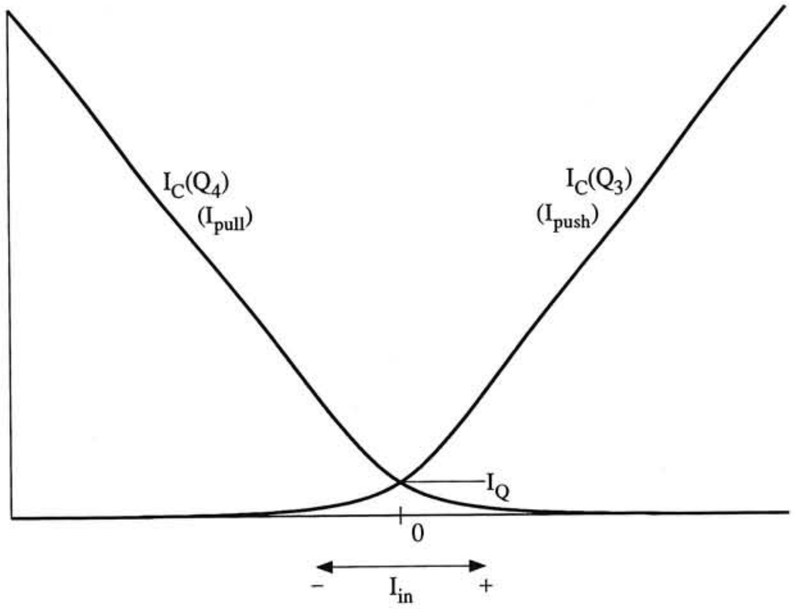


Figure 4.21: I_{push} and I_{pull} as a function of the input current

A convenient way to demonstrate the basic operation is to sketch the currents in Q_3 (I_{push}) and Q_4 (I_{pull}) as a function of the input current (Fig. 4.21).

A major advantage is the fact that neither of the output currents can ever become completely zero, which prevents excessive distortion by switching effects. Many variants of the traditional circuit with special specifications (e.g. for supply voltages below 2 V_{BE}) are found in literature. Some of them are true translinear circuits, others are not.

4.3.5 Miscellaneous translinear circuits

Apart from the translinear circuits described in the foregoing a lot of other special applications of the principle have been developed. These circuits provide special functions such as trigonometric functions, signal normalization, minimum and maximum functions. However, they lie beyond the scope of this chapter and we resort to referring to literature [3].

4.4 Systematic design methods for TL circuits

4.4.1 Introduction

Because TL circuits show common topological properties, they invite a systematic design approach. Seevinck [4] has carried out extensive research into the analysis and synthesis of TL circuits with bipolar transistors of the same polarity. The synthesis method is restricted to TL structures with less than 10 branches and one or two loops. Thus, all possible topologies of TL-circuits with the restrictions mentioned and with a number of prescribed transfer functions can be synthesized. However, it is surprising that nearly all fruitful and promising topologies found, have earlier been found with heuristic methods. But this is not true in all cases. As the methods are mainly based on network-theoretical and mathematical grounds, an extensive treatment lies beyond the scope of this book. Therefore we confine ourselves to a brief outline in Section 4.4.2 and refer to literature for details [4]. Further, an example of a useful TL circuit obtained by synthesis, that was not found earlier, will be shown.

4.4.2 The Seevinck synthesis method for bipolar semiconductor devices; an example of a systematically found TL circuit

The general aim is the design of TL networks realizing a prescribed (non)linear, time-invariant transfer function. The strategy shows some similarity with traditional synthesis methods for passive networks. The synthesis procedure can be divided into four general steps:

1. Approximation of the prescribed function by suitable algebraic formulations

2. Decomposition of the algebraic formulations found into forms suitable for TL realization

3. Realization of networks, based on topological properties of those TL networks, which fit with the forms found in 2.

4. Selection of the networks found as to their complexity, cost, stability, sensitivity to parameter tolerances, etc.

The parts 1 through 3 will briefly be explained now.

1: function approximation Only *algebraic* functions are suitable, of which rational functions need special attention, because they provide greater precision than polynomials of the same degree. Hence, non-algebraic functions need to be approximated by algebraic functions. As an example, a pretty accurate approximation of a cosine function is given below

$$\cos \pi X \approx \frac{(1-4X^2)(2-X^2)}{2+X^2} \ , \ |x| \leq 1 \tag{4.32}$$

2: function decomposition For synthesis purposes it is convenient to write the TL relation (4.6) in a slightly different form. Say that a TL loop has N elements (branches), numbered from 1 to N with branch currents I_1 through I_N, divided into odd and even ones, and with device areas A_1 through A_N, then (4.6) can be written as

$$\prod_{n=1}^{N/2} I_{2N} = \lambda \prod_{n=1}^{N/2} I_{2n-1} \text{ , where } \lambda = \prod_{n=1}^{N/2} \frac{A_{2n}}{A_{2n-1}} \tag{4.33}$$

Any TL network has one or more input currents $I_{i1,2,...}$ and output currents $I_{o1,2...}$ Further, every branch currents I_1 through I_N can be expressed in linear combinations of the input and output currents.

Generally, if the expressions of the branch currents into the input and output currents are called f_1 through f_N, application of (4.33) yields

$$f_1(I_{i1,2,...}I_{o1,2,...})f_3(I_{i1,2,...}I_{o1,2,...})\ldots = \lambda f_2(I_{i1,2,...}I_{o1,2,...})f_3(I_{i1,2,...}I_{o1,2,...})\ldots \tag{4.34}$$

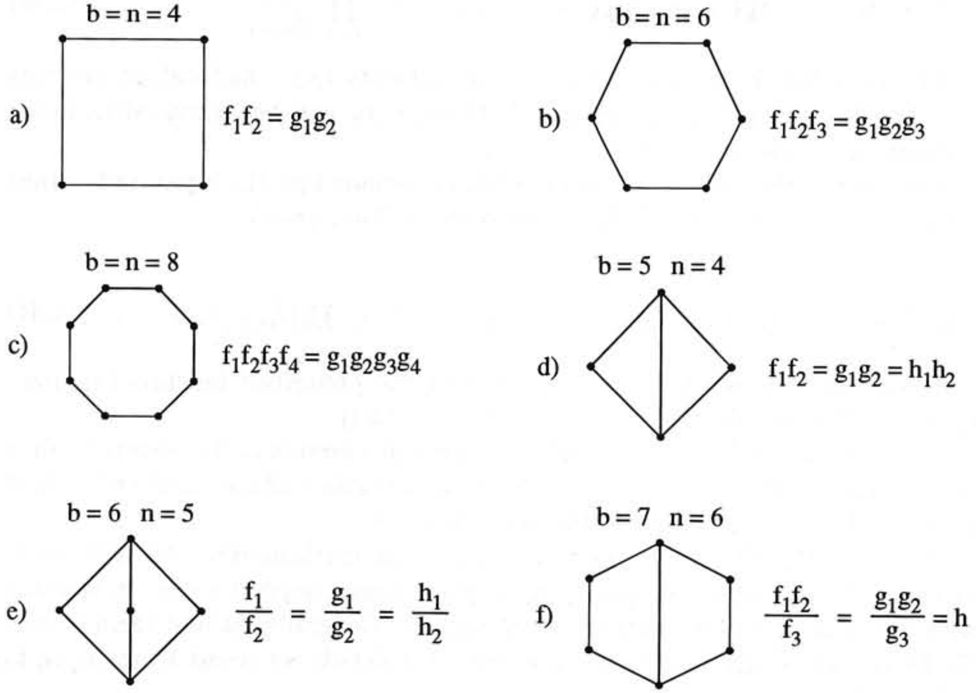
Hence, function decomposition means that the prescribed function (approximation) is "translated" into forms according to (4.34).

Note: As all functions f_1 through f_N represent currents in TL elements, they must remain positive for all (positive and negative) values of the input and output currents. This must be checked after decomposition.

Many decomposition techniques are known in mathematics. Suitable techniques for TL synthesis are those using *explicit forms; implicit forms; parametric forms; rational functions; continual fractions, etc.* Generally, each of them is suitable for a class of function approximations. For details we resort to referring to literature [4].

3: network realization techniques An arbitrary TL network always contains one or more (interwoven) TL loops with minimally four branches. If the branch currents and node voltages are left out of consideration and if, besides, every TL element is symbolized by a line, the result is the so-called *undirected graph* of the

TL network. Every graph represents a class of TL networks. Of course, the number of branches (= TL elements) is theoretically unlimited. However, due to practical parameter tolerances it has been shown to be senseless to construct TL networks with more than 9 branches and/or more than 2 loops. This limitation results in maximally 6 different graphs. Any graph can more precisely be characterized by numbering its nodes and choosing the direction of the branch currents. Then every graph has a corresponding *node-branch incidence matrix (the T matrix).* The total number of different T matrices corresponding to the 6 graphs amounts to 26. To date, the connections of the in- and output currents and the values and connections of biasing currents have not yet been chosen. Hence, it will be clear that any T matrix generally results in a great number of possible TL networks. Checking them all would be possible, but this immense job would be entirely a matter of analysis, and give hardly any insight. To make a real synthesis of TL circuits feasible, the possible general function structure of the relations between the branch currents that can be realized by any graph, has to be investigated.



b = n = 4
a)
$f_1f_2 = g_1g_2$
b = n = 6
b)
$f_1f_2f_3 = g_1g_2g_3$
b = n = 8
c)
$f_1f_2f_3f_4 = g_1g_2g_3g_4$
b = 5 n = 4
d)
$f_1f_2 = g_1g_2 = h_1h_2$
b = 6 n = 5
e)
$\frac{f_1}{f_2} = \frac{g_1}{g_2} = \frac{h_1}{h_2}$
b = 7 n = 6
f)
$\frac{f_1f_2}{f_3} = \frac{g_1g_2}{g_3} = h$

Figure 4.22: The six undirected graphs. b and n are the numbers of branches and nodes

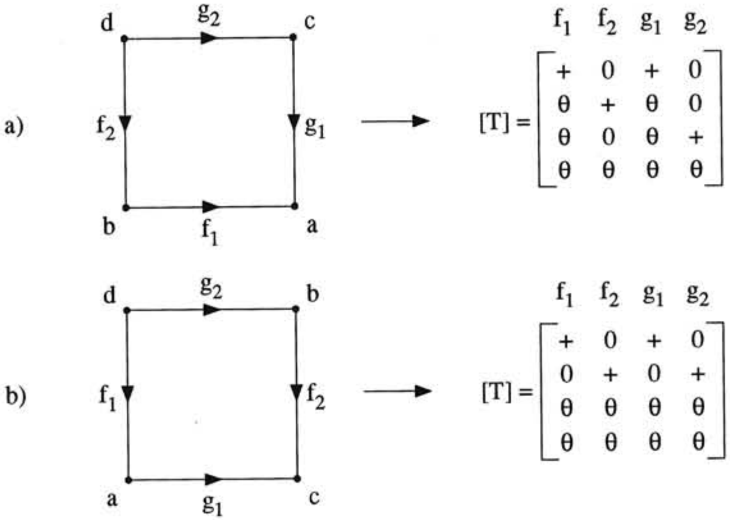
Fig. 4.22 shows the 6 (undirected) graphs with their appropriate general function structures. Now the synthesis procedure is as follows. First, the (approximation of the) desired function is decomposed in one or more ways, so that the results fit with one (or more) of the general function structures shown in Fig. 4.22. Second, all possible T matrices are derived from the (directed) graphs. Third, all possible TL networks are derived from the T matrices. Finally, the resulting networks are checked and selected on feasibility and quality. The complete procedure will be demonstrated with a simple example: Say, a TL two-quadrant divider with transfer $z = x/y$, with $|x| < y$ must be made. The function can be decomposed into

$$z + 1 = \frac{x+y}{y} \text{ or } y(z+1) = x + y \tag{4.35}$$

or into

$$\frac{1+z}{1-z} = \frac{y+x}{y-x} \text{ or } (1+z)(y-x) = (1-z)(y+x) \tag{4.36}$$

Both functions fit with the general form $f_1 f_2 = g_1 g_2$, where all functions remain positive. From Fig. 4.22 the first graph is selected. After directioning, this graph results in two possible T matrices [4] (see Fig. 4.23).



$$[T] = \begin{array}{c} \begin{array}{cccc} f_1 & f_2 & g_1 & g_2 \end{array} \\ \begin{bmatrix} + & 0 & + & 0 \\ \theta & + & \theta & 0 \\ \theta & 0 & \theta & + \\ \theta & \theta & \theta & \theta \end{bmatrix} \end{array}$$

$$[T] = \begin{array}{c} \begin{array}{cccc} f_1 & f_2 & g_1 & g_2 \end{array} \\ \begin{bmatrix} + & 0 & + & 0 \\ 0 & + & 0 & + \\ \theta & \theta & \theta & \theta \\ \theta & \theta & \theta & \theta \end{bmatrix} \end{array}$$

Figure 4.23: The two possible T matrices

According to the decomposed functions (4.35) and (4.36), the two TL networks shown in Fig. 4.24 appear to be feasible. The left one follows from the T matrix of the left part of Fig. 4.23 together with (4.35) whereas the right one follows from the T matrix of the right part of Fig. 4.23 together with (4.36) (Note that the right circuit low-voltage with symmetrical outputs, whereas the left circuit is not).

It will be clear that with all 26 T matrices the realization of numerous different TL networks providing many (approximated) transfers is feasible [4].

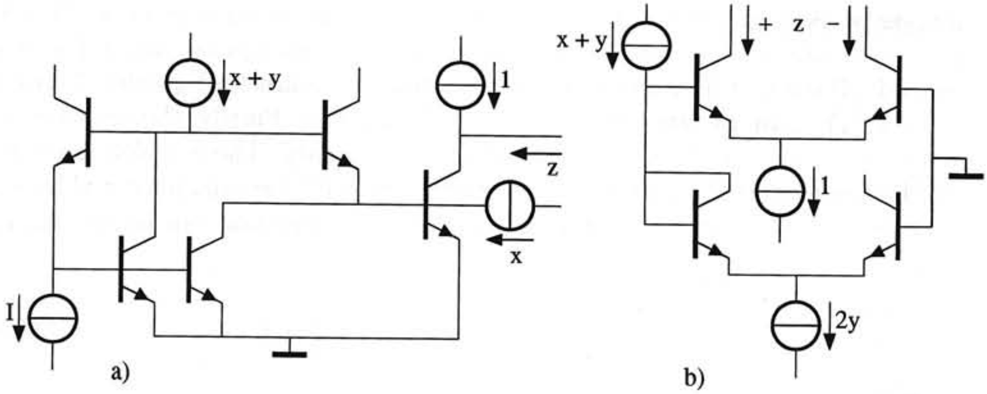


Figure 4.24: Two possible TL-circuits for a two quadrant divider

4.5 Recent and future developments

Most developments of TL circuits took place from 1968 to, say, 1988. The resulting products were ICs suitable for one or more particular signal processing functions, e.g. four-quadrant multiplying. Much research was carried out to improve their accuracy and other qualities. To date they are/were mainly used as building blocks in electronic systems containing many discrete ICs and other discrete components. Recently a revival of the interest in TL circuits has grown. However, the application areas are different now. First, TL circuits are often used together with other circuits on one chip. This often makes the designs of the available building blocks useless. Some of their characteristics exceed the demands, e.g. accuracy, whereas other characteristics are not good enough, i.e. they have too large chip area or they need too much supply voltage/power. Second, TL circuits are recently used for new purposes such as neural networks, which asks for classes of TL circuits with extremely large circuit density and which have entirely different characteristics than the available building blocks. Summarizing, it may be stated that there exists a need for new generations of TL circuits for *(extremely) low-voltage/low-power applications* and with *minimal chip occupation*. Recently a research program, where the possibilities of using the backgate of MOSTs in weak inversion for true translinear circuits are investigated, has been started, i.e. the MOST is considered as a four terminal device, where the backgate is intentionally used as a signal electrode. In the Appendix one of the first results is reported. A brand-new branch of research in translinear circuits concerns their application for dynamic signal processing, e.g. in filters.

References

[1] B. Gilbert, "Translinear circuits: A proposed classification", Elec. Letters, 11(1):14-16, January 1975.

[2] E.H. Nordholt, "Design of high-performance negative feedback amplifiers", Elsevier: Amsterdam, 1983.

[3] B. Gilbert, "Current-mode circuits from a translinear viewpoint: A tutorial", Ch. 2 of "Analogue IC design: the current-mode approach", Peter Peregrinus Ltd. London 1990.

[4] E. Seevink, "Analysis and synthesis of translinear integrated circuits", Elsevier: Amsterdam, 1988

Appendix: Translinear sin(x)-circuit in MOS technology using the back gate

J. Mulder, A.C. van der Woerd, W.A. Serdijn, A.H.M. van Roermund

Abstract

Though the MOS transistor is a four-terminal device, it is most often regarded as being a three-terminal device. Therefore, many possible MOS circuits are overlooked. In this paper, the four-terminal point of view is elaborated with respect to MOS weak inversion translinear circuits. It is shown that, by using the back gate, translinear networks can be derived which cannot be realized with bipolar transistors. These networks increase the possibilities offered by translinear technology. A sin(x)-circuit, which is one of the possible applications of the new network, was measured. The circuit can operate at supply voltages of less than 2 V and with a total bias current of only 14 nA.

I. Introduction

The first translinear circuits, published in '68 by Gilbert [1], were designed using bipolar transistors. However, MOS transistors in weak inversion are also suitable for this type of circuit because of the almost exponential relation between the gate-source voltage and the drain current in this region [2]. In contrast with the bipolar transistor, the MOST is a four-terminal device. In subthreshold, the relation between the bulk-source voltage and the drain current is also exponential. A sufficiently accurate model for the drain current of a MOST in saturation is given by [3]:

$$I_{DS} = I_0 \mathrm{e}^{V_{GS}/\kappa U_T} \mathrm{e}^{V_{BS}/\eta U_T} \tag{4.1}$$

where I_0 is the zero-bias current, V_{GS} and V_{BS} are the gate-source and bulk-source voltage, $U_T = kT/q$ is the thermal voltage and κ and η are the inverses of the subthreshold slopes, which are constant in this model.

A simple way to design a MOS translinear circuit is to translate a bipolar circuit directly to its MOS equivalent, replacing the base-emitter junctions by gate-source voltages and connecting the substrate terminal of each MOS transistor to its source. Using this approach, the functionality of the substrate terminal as a second gate, or back gate, is not recognized and therefore a class of new circuits is ruled out in advance. As shown in this paper, the use of the back gate enables us to design translinear circuits that are not possible when using bipolar transistors. As an example, a sin(x)-circuit is presented. Measurements of a breadboard version of the circuit are shown.

II. Translinear topology

The new translinear circuit topology is depicted in Fig. 4.1. The circuitry necessary to bias the MOSTs at the proper drain currents is not shown. For an NMOS implementation, as shown in Fig. 4.1, a double well process will be necessary. Of course, if the circuit is implemented in PMOS, only n-wells will be needed.

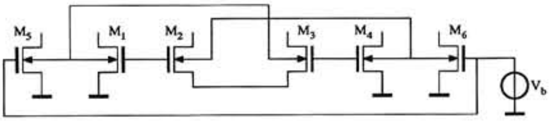


Figure 4.1: Topology described by $\frac{I_1 I_3}{I_2 I_4} = \frac{I_5^2}{I_6^2}$

The circuit topology consists of a four-transistor gate-source loop in the up-down topology and two additional MOSTs biased at the same gate voltage. Of course, MOSTs with different back gate voltages have to be integrated in separate wells. Thus, M_1, M_3 and M_5 in the first well, and M_2, M_4 and M_6 in the second well. The back gates of M_1 and M_3 are connected together and the same applies for M_2 and M_4. The back gate voltages of M_1 and M_2 are determined by connecting their back gates to the back gates of M_5 and M_6. These two transistors have to be biased at the same gate voltage to obtain a theoretically process- and temperature-independent transfer function. Using the simple drain current model (4.1), the topology is described by an equation containing two squared currents:

$$\frac{I_1 I_3}{I_2 I_4} = \frac{I_5^2}{I_6^2} \tag{4.2}$$

The two squared currents result from the connection of the back gates of M_1 and M_3 and of M_2 and M_4. Because of this connection, the back gate voltages of M_1 and M_3 and of M_2 and M_4 are added, resulting in two factors 2. These factors 2 are the two exponents on the right-hand side of (4.2).

Equation structure (4.2) is different from the four different equations that can be realized with bipolar translinear networks [4]. Thus, this topology increases the number of possible translinear solutions for the realization of a given function. As this equation structure is more complex than the four mentioned 'bipolar' equation structures, in some cases, a higher functional density and thus area-efficiency can be obtained.

The topology shown in Fig. 4.1 can also be regarded in another way; the circuit consists of two loops of gate-bulk voltages. The first loop is formed by M_1, M_2, M_6 and M_5. The second by M_3, M_4, M_6 and M_5. The sources of M_1, M_4, M_5 and M_6 are connected to ground. The sources of M_2 and M_3 are tied together. Since no gate-source voltages are connected in series, the circuit is suitable for low-voltage applications.

III. Sin(x) circuit

As an example of the use of the new topology, a differential sin(x)-circuit was designed. Since the transfer function of a translinear circuit is always a rational function, an approximation for the sine function has to be used. According to [5], the sine function can be approximated by:

$$z = \sin \pi x \approx \frac{x - x^3}{1 + x^2} \tag{4.3}$$

where x and z represent the normalized input and output current, respectively. Another way of writing this approximation is the implicit decomposition [4]:

$$\frac{1 + z + x}{1 - z - x} = \frac{(1 + x)^2}{(1 - x)^2} \tag{4.4}$$

This decomposition can easily be fitted on equation structure (4.2) by choosing $I_2 = I_{bias} - I_{out} - I_{in}$, $I_3 = I_{bias} + I_{out} + I_{in}$, $I_5 = I_{bias} + I_{in}$, $I_6 = I_{bias} - I_{in}$ and $I_1 = I_4$. The sine shaped output current is obtained from $I_3 - I_2 - 2I_{in} = 2I_{out}$.

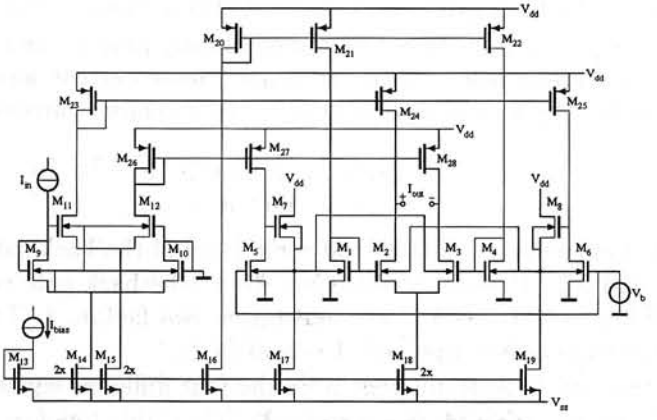


Figure 4.2: Sin(x)-circuit

The complete circuit is depicted in Fig. 4.2. M_7 and M_8 are two simple floating voltage sources, which are used to keep M_5 and M_6 in saturation for bulk voltages of less than about 100 mV. Since the circuit is differential, a gain cell M_9 to M_{12} [1] is used to convert the input signal into a differential signal. Current mirrors are used to supply the currents to the actual sin(x)-circuit.

The applications of the general topology shown in Fig. 4.1 are not restricted to the example treated in this paper. Many other functions will fit on the topology, which in fact is the main strength of translinear technology.

IV. Measurement results

A trivial application of (4.2) is the construction of a $\sqrt{x}$-circuit. To verify the new equation structure (4.2), a breadboard verion of the $\sqrt{x}$-circuit was measured. The drain currents through M_2, M_3, M_4 and M_6, shown in Fig. 4.1, are all biased at 1 nA. The drain currents of M_1 and M_5 are the input and output current, respectively. The gates of M_5 and M_6 are biased at 550 mV. The aspect ratios of the used NMOSTs are 108/7 μm/μm.

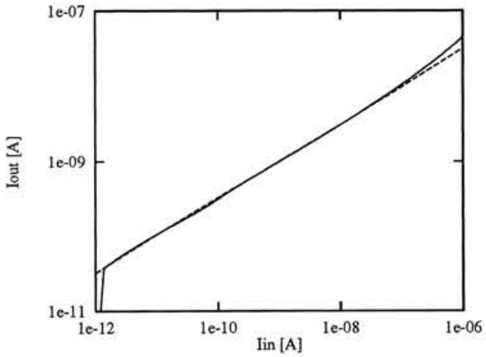


Figure 4.3: Output current of the square root circuit

Measurements were performed using an HP4142B Modular DC Source / Monitor. In Fig. 4.3, the measured output current is compared with the theoretical curve. Clearly, the output current is proportional to the square root of the input current. The large errors at low and high values of the input current are caused by leakage currents of the measurement set-up and by the transition into the moderate inversion region, respectively. The main cause of error for intermediate current values is mismatch; the mismatch was quite large due to the breadboard realization. The average mismatch between the drain currents of two transistors at the same gate-source voltage was about 9%.

Next the sin(x)-circuit, shown in Fig. 4.2, was measured. The measured output current is shown in Fig. 4.4. The gates of M_5 and M_6 are biased at 350 mV. The supply voltages V_{dd} and V_{ss} are ±1V and can even be lower, in principle. The bias current I_{bias} is 1nA, resulting in a total bias current of only 14 nA. The input current I_{in} ranges from 0 to 2 nA. The drains of M_2 and M_3 are loaded by two 500 mV voltage sources. Despite the rather large mismatch, due to the breadboard realization, which causes offset, asymmetry, amplitude, phase and frequency errors, the result is quite reasonable, as is shown by the comparison of the measured output current with a fitted sine function, see Fig. 4.4.

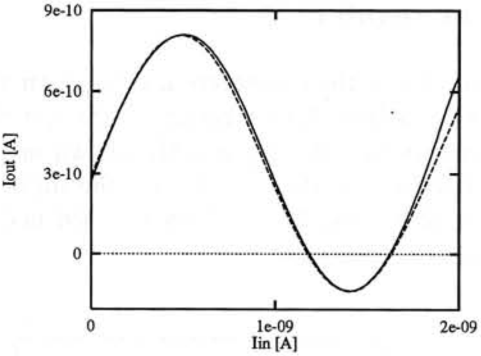


Figure 4.4: Measured output current (—) of the sin(x)-circuit and a fitted sine function (– –)

V. Conclusions

Regarding the MOS transistor as a four-terminal device with a front and a back gate, a new translinear circuit topology was derived. This equation structure increases the number of possible designs for a certain function to be realized in translinear technology, and might result in more area-efficient implementations. As an example of the new topology, a sin(x)-circuit was designed. The circuit operates at supply voltages of less than 2 V, with a total bias current of only 14 nA. Measurements were performed which verify the theory, although they suffer from rather large mismatch of the MOSTs due to the breadboard realization.

References

[1] B. Gilbert. A new wide-band amplifier technique. *IEEE Jour. of Solid State Circ.*, 3(4):353–365, December 1968.

[2] Y.P. Tsividis. *Operation and modeling of the MOS transistor*. McGraw-Hill, 1987.

[3] A. Pavasović. *Subthreshold region MOSFET mismatch analysis and modeling for analog VLSI systems*. PhD thesis, John Hopkins University, Baltimore, Maryland, 1991.

[4] E. Seevinck. *Analysis and synthesis of translinear integrated circuits*. Elsevier, 1988.

[5] B. Gilbert. Translinear circuits: A proposed classification. *Elec. Letters*, 11(1):14–16, January 1975.

Chapter 5

Integrable DC sources and references

Arie van Staveren

5.1 Introduction

Electronic systems can be seen as an implementation of mathematical functions. A system may contain all kinds of blocks, for example, integrators, multipliers and constants.

This chapter discusses system blocks, which are integrable, for generating a:

- voltage constant,
- current constant.

These constants are widely used in electronic design. For example, a current constant is used to determine the collector bias current of a transistor. In this case, the absolute accuracy is not that important. Mostly, a relatively large variation as a function of temperature, time or other parameter is allowed.

Another example is a voltage constant used in a voltmeter. An unknown voltage is compared indirectly with the voltage constant to determine its value. The absolute value of the constant is of prime importance. The accuracy of the measurement cannot be better than the accuracy of the voltage constant.

These two examples show two specific types of application of constants:

- The constant used as a source, i.e. a voltage or current *source*. The absolute value is not particularly important. The value may change within a certain region.
- The constant used as a reference, i.e. a voltage or current *reference*. The absolute value of the constant is of prime importance.

The application determines when the implementation of a constant is called a reference or a source. Both names are used in this chapter. When general theory of constants is treated, the most general name, the source, is used. When more specific implementations are treated, the most commonly used name is employed.

The next section of this chapter starts with the description of the ideal voltage and current constants. Using this description, implementations of voltage constants are given, from very simple (resistive divider) to more complicated circuits (bandgap reference). The non-ideal effects, like the finite current-gain factor, Early effect, etc. are discussed. The current constant is treated after the voltage constant, because the current source is often derived from a voltage source via a (trans)conductance. Implementations and non-ideal effects are discussed.

5.2 The ideal voltage and current constants

Voltage and current constants are generated, respectively, by a voltage source and a current source. The output signal of ideal sources is independent of the load, temperature and all other kinds of environmental disturbances. Further, the output signal is not contaminated by noise.

5.2.1 The ideal voltage source

In figure 5.1, the output signal versus the load current of the ideal voltage source is depicted in the V-I plane. As can be seen, the output voltage V_{ref} is independent of the load current I_{load}. As the output impedance of a voltage source is defined as the ratio of output-voltage variation and load-current variation, the output impedance r_{out} of the ideal voltage source equals zero:

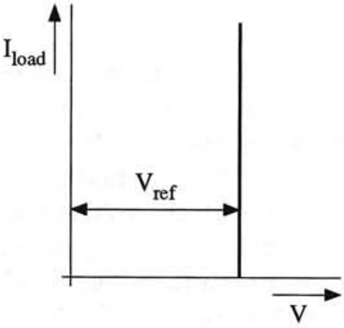


Figure 5.1: The output voltage V_{ref} versus the load current I_{load}

$$r_{\text{out}} = \frac{dV_{\text{ref}}}{dI_{\text{load}}} = 0\Omega. \tag{5.1}$$

The ideal output voltage is not influenced by a change in the ambient temperature. Its temperature coefficient is zero:

$$\alpha_T = \frac{dV_{\text{ref}}}{dT} = 0\text{V/K}. \tag{5.2}$$

Finally, all the power supplied has to be concentrated at dc. The signal-to-noise ratio of the output voltage is infinite. Sometimes it is more convenient to talk about the absolute value of the output noise. For the ideal case the power-density spectrum S_v of the noise voltage at the output equals:

$$S_v = 0\text{V}^2/\text{Hz}. \tag{5.3}$$

5.2.2 The ideal current source

In figure 5.2, the output current I_{ref} of the ideal current source is depicted in the I-V plane as a function of the load voltage. The output current is independent of this voltage. Therefore the output impedance r_{out} of the ideal current source is infinite:

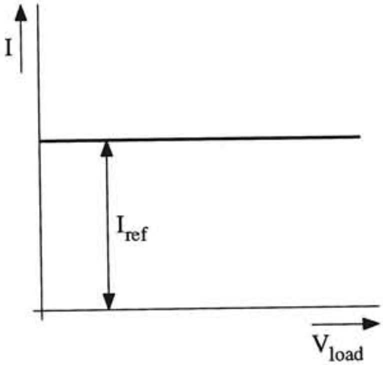


Figure 5.2: The output current I_{ref} as a function of the load voltage V_{load}

$$r_{\text{out}} = \frac{dV_{\text{load}}}{dI_{\text{out}}} = \infty\Omega. \tag{5.4}$$

Just like the ideal voltage source, the temperature coefficient of the output current and the power-density spectrum S_i of the noise current at the output are zero:

$$\alpha_T = 0\text{A/K}, \tag{5.5}$$
$$S_i = 0\text{A}^2/\text{Hz}. \tag{5.6}$$

Integrable implementations of the ideal voltage and current source are discussed in the following sections.

5.3 Implementations of the voltage source

As discussed in the previous section, an ideal voltage source has specific characteristics. Summarized:

- output impedance of 0Ω,
- temperature independent,
- noise free.

When practical implementations are made, the voltage source itself needs a power supply. The output voltage of the source must be independent of this power supply. Power-supply variations are not allowed to penetrate to the output of the voltage source. Thus the power supply only supplies the dc bias current. The figure of merit for this quality aspect is the Power Supply Rejection Ratio or PSRR for short. The PSRR is defined as:

$$PSRR \triangleq \frac{dV_{\text{power}}}{dV_{\text{ref}}} \tag{5.7}$$

with V_{power} the supply voltage. For a voltage source approximating the ideal source it holds that

- The Power Supply Rejection Ratio has to be infinite.

Depending on the application, one or more of the four mentioned constraints are of prime importance.

The output voltage of a realistic voltage source is depicted in figure 5.3. The source behaves as a voltage source when the bias current is above the threshold current I_{th}. Thus practical implementations pose additional constraints.

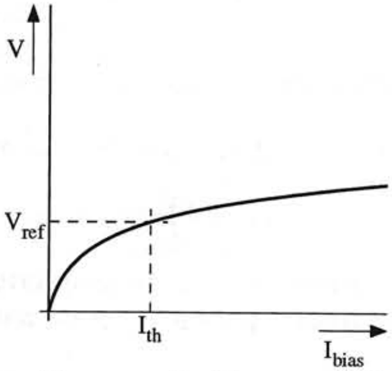


Figure 5.3: The output voltage V_{ref} versus the bias current I_{bias} of a realistic voltage source

Several implementations of voltage sources are discussed in the next section.

5.3.1 The resistive divider

The schematic of the voltage source implemented by a resistive divider is depicted in figure 5.4. The output voltage V_{ref} is a fraction of the power-supply voltage V_{cc}:

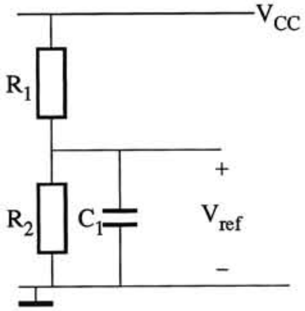


Figure 5.4: A voltage source implemented by a resistive divider

$$V_{\text{ref}} = \frac{R_2}{R_1 + R_2} V_{\text{CC}}. \tag{5.8}$$

The accuracy of this source is determined by the accuracy of the power-supply voltage and the matching of the two resistors.

The output impedance, r_{out}, of this source equals the parallel connection of the two resistors:

$$r_{\text{out}} = \frac{R_1 R_2}{R_1 + R_2}. \tag{5.9}$$

To obtain a low output impedance with this source, low resistances have to be used. This results in a high current consumption. This relation becomes more clear when the total supply current I_{bias} is expressed as a function of the output impedance. The following relation is found:

$$I_{\text{bias}} = \frac{V_{\text{CC}}}{R_1 + R_2} = \frac{1}{r_{\text{out}}} \left(1 - \frac{V_{\text{ref}}}{V_{\text{CC}}}\right) V_{\text{ref}}. \tag{5.10}$$

Thus for a given output voltage and power-supply voltage, the bias current is inversely proportional to the output impedance.

When the output impedance is important for higher frequencies only, the impedance can be made low with a capacitor. This is illustrated in figure 5.4 with capacitor C_1. Now C_1 determines the output impedance for relatively high frequencies. This capacitor may be too large for integration. In this case an additional pin is required to be able to keep the capacitor outside the chip.

The output noise is determined by the thermal noise of the parallel connection of the two resistors. The power-density spectrum, S_v, of the noise voltage equals:

$$S_v = 4kT(R_1 \parallel R_2). \tag{5.11}$$

Again, for a low-noise behavior, low resistances and thus a high current consumption are required. With capacitor C_1 this power-density spectrum can be reduced for the relatively high frequencies.

The PSRR is determined by the resistors and is given by:

$$PSRR = 1 + \frac{R_1}{R_2}. \tag{5.12}$$

To obtain a high power supply rejection ratio, the ratio R_1 and R_2 has to be large. However, for a given output voltage, this ratio is fixed. A solution is to decouple R_2 by using a sufficiently large capacitor. This can also be capacitor C_1 in figure 5.4.

The temperature behavior of this source is determined by the temperature stability of the resistor ratio and the temperature dependency of the supply voltage.

5.3.2 The non-linear divider

To improve the performance of the source, a non-linear impedance can be used for R_2. The principle is depicted in figure 5.5. The non-linear function is the I-V characteristic of the non-linear device (NL), the linear one is of the resistor (R). The output voltage is given by the intersection point of the two functions.

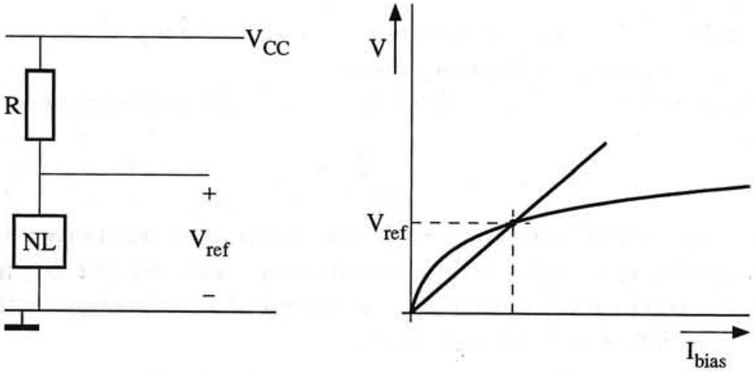


Figure 5.5: A voltage source using a non-linear device

With this non-linear device, the large-signal behavior (the generation of an output voltage) and the small-signal behavior (a low output impedance) are different. The output impedance of the source is approximately the small-signal impedance of the non-linear device. This can be much lower than for the resistive divider.

The noise behavior of this source with respect to the linear divider with the same current consumption generally improves.

As seen in the previous section, the PSRR [see equation (5.12)] increases when R_2 decreases. For this source, R_2 is replaced by the small-signal impedance of the non-linear device. The PSRR can be considerably higher.

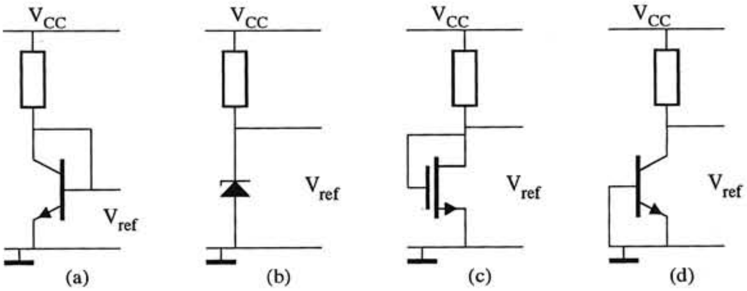


Figure 5.6: Four non-linear devices

Various devices can be used for the non-linear part of the divider. Four possibilities are depicted in figure 5.6. The devices are:

a. diode-connected bipolar transistors,

b. diodes at reverse Breakdown,

c. diode-connected normally-off FETs,

d. bipolar transistors used at punch-through.

A diode-connected bipolar transistor

In figure 5.6a, the diode-connected bipolar transistor is used as the non-linear device. The output voltage equals the base-emitter voltage of the transistor and is in the range of 0.5V to 0.8V, depending on the collector-current density. For higher values more junctions in series have to be used. Lower voltages can be obtained with a Schottky junction, for instance, with a junction voltage of approximately 0.2V. The output impedance r_{out} is approximately:

$$r_{\text{out}} \approx \frac{kT}{qI_{\text{bias}}} \tag{5.13}$$

with I_{bias} the collector bias current of the transistor. The higher the current the lower the output impedance. The temperature dependency of this source is the temperature dependency of the junction voltage and is in the order of a few mV/K.

The noise of the source equals the noise generated by the transistor, when it is assumed that the noise of the resistor is negligible. The power-density spectrum S_v of the noise voltage equals:

$$S_v = 4kT(r_b + \frac{1}{2g_m}) \tag{5.14}$$

with r_b the base resistance and g_m the transconductance of the transistor. An improvement of the noise performance is obtained when higher bias currents (the g_m reduces) or larger transistors (the base resistance reduces) are used.

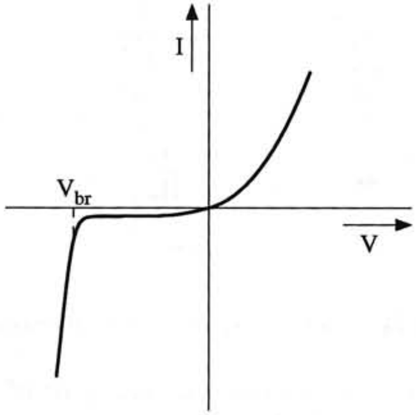


Figure 5.7: The V-I characteristic of a Zener diode

A diode at reverse breakdown

A voltage source can also be made of a reverse-biased diode at breakdown. Zener diodes are optimized for use in this mode. The V-I characteristic of a Zener diode is depicted in figure 5.7. The forward behavior is comparable to that of a normal pn junction. But when the Zener diode is biased in reverse mode and the voltage is increased slowly, at a specific voltage the currents suddenly starts to increase very rapidly. This specific voltage is called the reverse-breakdown voltage, V_{br}. Beyond this voltage the Zener diode behaves like a voltage source (c.f. figure 5.1). For normal diodes this reverse breakdown can be destructive.

The reverse breakdown is due to two distinct mechanisms, *avalanche multiplication*, which causes an avalanche breakdown and the *Zener effect*, which causes a Zener breakdown. Although diodes are optimized such that one of the two mechanisms is dominant, both types of diodes are called Zener diodes.

Avalanche breakdown

Avalanche breakdown occurs in the presence of high electric fields in relatively wide regions. The carriers are accelerated sufficiently to become able to ionize atoms. The newly created carriers are accelerated and the also ionize atoms. An avalanche of carriers arises and the current increases very rapidly. This effect is more apparent in lightly-doped materials and is proportional to the electric field strength. In lightly-doped materials, the carriers are able to travel a relatively large distance without collisions. The chance that they obtain enough energy for ionizing other atoms increases. For higher electric field strengths, the free path length, necessary for ionizing other atoms, decreases, more ionizations per unit length occur and the current increases. Avalanche breakdown is the dominant

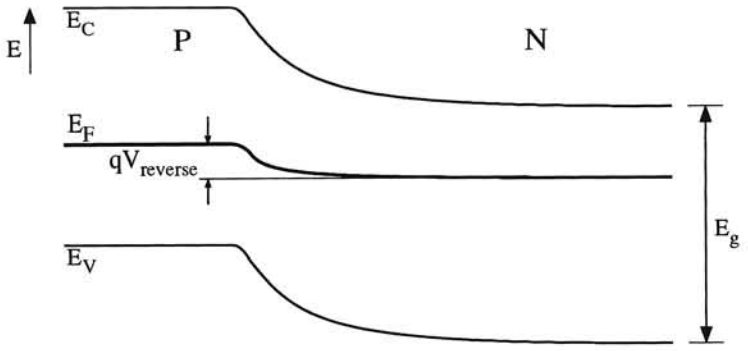


Figure 5.8: The band diagram of a slightly reverse-biased junction

breakdown effect in junctions with a breakdown voltage higher than 6V.

When there is avalanche breakdown, the junction is overwhelmed by high-energetic carriers and the junction may be damaged. This results in an increase of 1/f and the non-ideal currents. The increase of the non-ideal currents is the cause of the deterioration in the low-current behavior. Thus, in normal junctions, avalanche breakdown has to be avoided.

Zener breakdown

In slightly reverse-biased junctions, the energy bands of a junction are as depicted in figure 5.8. Between the valence band (E_V) and conduction band (E_C) the energy gap E_g exists. A carrier has to gain an amount of energy equal to or more than this E_g to be able to reach the other band. When the diode is reverse biased and the reverse voltage is increased, the bottom of the conduction band and the top of the valence band of, respectively, the n and p material reaches the same energy level. When the distance W (see figure 5.9) between the bands is smaller than a critical value, the carriers can tunnel directly from the valence band in the p material to the conduction band in the n material, without the help of other carriers. A further increase in the reverse voltage leads to an overlap of the two bands (see figure 5.9) and the potential-barrier width, W, becomes smaller and smaller. Due to this reduction of barrier width, the tunneling probability increases and the current increases even further. For higher-doped junctions, the depletion region is less wide and results in a lower Zener breakdown voltage. Since tunneling can start only when the bands are at equal levels, Zener breakdown occurs more abruptly than avalanche breakdown. Pure Zener breakdown occurs in highly doped junctions with a reverse-breakdown voltage lower than 5V.

The temperature dependency of the diode at reverse breakdown

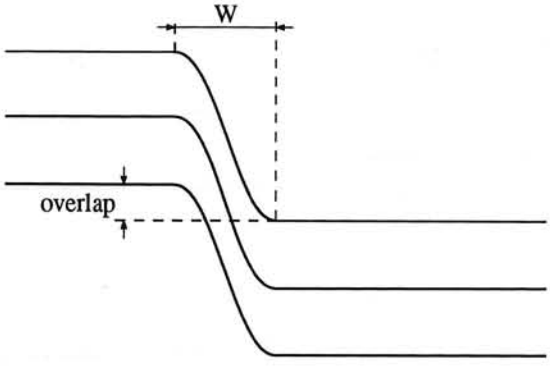


Figure 5.9: The band diagram when the conduction band and the valence band have an overlap

The avalanche effect depends on the free path length of the carriers. When the temperature rises, lattice vibrations increase and the carriers are hindered more. The mobility decreases and a higher electric field is needed to reach breakdown. Thus, the avalanche breakdown voltage has a positive temperature coefficient. However, the Zener breakdown is strongly dependent on the barrier width. For higher temperatures, the barrier width reduces and the Zener breakdown occurs at a lower voltage. Thus, the Zener breakdown voltage has a negative temperature coefficient.

When the temperature dependency of the breakdown voltage is too high, temperature compensation has to be carries out. This can be accomplished in two ways. Firstly, the diode can be constructed in such a way that the Zener effect and the avalanche multiplication are both equally important. As the temperature behavior of these two effects are opposite, a zero temperature coefficient can be obtained. A second method is to use a forward-biased junction. The temperature coefficient of a forward-biased junction is negative (this is discussed later) and about -2mV/K. The temperature coefficient of the avalanche breakdown voltage is about +2mV/K. With a series connection of a reverse-biased Zener diode at avalanche breakdown and a forward-biased junction, the temperature coefficient can be nullified. The voltage necessary is approximately 7V, being about 6V for the Zener diode and about 0.8V for the forward-biased diode.

Zener diodes on chip

The diodes at reverse breakdown frequently used on chip are in the range below 5V. Thus, the dominant effect is the Zener effect. In IC technology two types of Zener diodes can be realized. Firstly, a Zener diode made of a junction at the surface of the chip. Because this diode is mainly located at the surface, surface

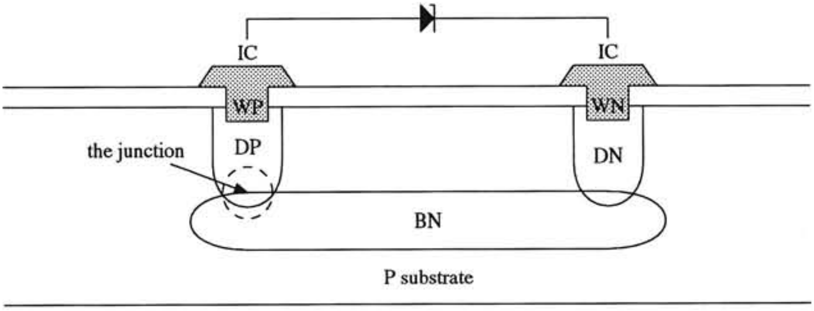


Figure 5.10: The realization of a Zener diode in the bulk of a chip

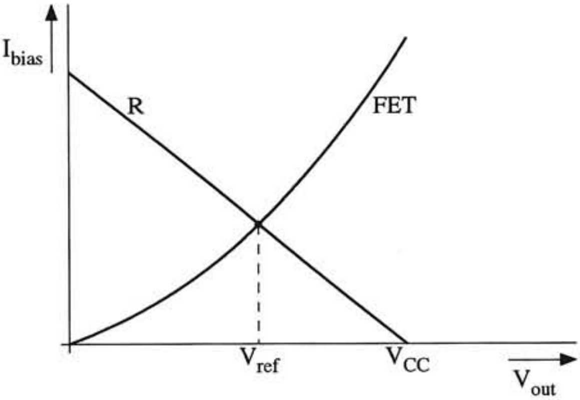


Figure 5.11: The output voltage of the voltage source using a FET

effects like 1/f noise have a greater influence on the behavior of the diode. Secondly, the Zener diode can be made in the bulk of the chip, see figure 5.10 The diode is formed by the buried N layer (BN) and the deep P diffusion (DP the isolation diffusion). Both are highly doped and Zener breakdown is likely to occur. In the design manual, it can be seen that the corresponding breakdown voltages are low. This diode is completely surrounded by bulk material and, consequently, the behavior is less hampered by surface defects.

Diode-connected normally-off FETs

When a normally-off FET is used as the non-linear component in the divider shown in figure 5.5, the output voltage is determined by the intersection point of the V_{GS} versus I_R and V_R versus I_R (see figure 5.11).

The output impedance, r_{out}, of this source equals approximately:

$$r_{\text{out}} = \frac{1}{g_m} \tag{5.15}$$

with g_m the transconductance of the FET. For lower output impedances, more current is required. The noise of this source is due to the thermal noise of the resistor and the drain noise of the FET. The PSRR equals:

$$PSRR = \frac{1}{1 + g_m R}. \tag{5.16}$$

A transistor used at punch-through

When a junction is biased in reverse mode, the depletion layer becomes wider for higher reverse voltages. In a transistor, the base-collector junction is mostly reverse biased. The depletion layer of the base-collector junction reduces the effective base width (modeled by the forward Early voltage). When the reverse voltage is increased such that the depletion layer of the base-collector junction touches the depletion layer of the base-emitter junction, the effective base width is reduced to zero. An electric field now exists across this depleted area and transports every carrier that enters the region to the other side. Similar to the current through a collector-base depletion layer, the field cannot influence the number of carriers transported. The current is determined by the supply of carriers at the depletion layer boundaries. Since the emitter is highly doped, the current can become very large and an external current-limiting resistor has to be connected in series with the collector lead (see figure 5.6). For an increasing current, the voltage across the resistor increases, consequently, the voltage across the transistor decreases. At the biasing point, the voltage is such that the base-emitter and base-collector depletion regions just touch each other.

The output voltage of this circuit is indirectly determined by the number of available carriers. A small increase in the reverse base-collector voltage results in a very large increase in the output current. Thus, the output impedance of this source is very low.

5.3.3 Diode-connected transistors in forward mode

In this section, a base-emitter junction is used as an element with a very well-known I-V relation and temperature behavior. The voltage reference obtained can be very accurate with respect to output voltage and temperature behavior. The relation between the current and the voltage of a diode-connected transistor is firmly stated by physical relations. The I-V relation is given by:

$$I_C = I_S \left[\exp\left(\frac{qV_{BE}}{kT} \right) - 1 \right] \tag{5.17}$$

in which I_C is the collector bias current, I_S the collector saturation current, q the electron charge ($1.6 \cdot 10^{-19}$C), T the absolute temperature and k Boltzmanns constant ($1.38 \cdot 10^{-23}$J/K). The -1 term is negligible. Even for very small base-emitter voltages, the exponential term is already much larger.

The base-emitter voltage is the parameter of interest, so relation (5.17) is rewritten to make V_{BE} explicit. The expression for V_{BE} is:

$$V_{BE} = \frac{kT}{q} \ln\left(\frac{I_C}{I_S}\right). \tag{5.18}$$

With $\frac{kT}{q}$ the thermal voltage. At room temperature ($\approx$ 300K) the thermal voltage equals:

$$\frac{kT}{q} \approx 26\text{mV}. \tag{5.19}$$

Because of the logarithm, a multiplicative change in the collector current becomes an additive change in the base-emitter voltage. For example, when the collector current is increased by a factor 10, the base-emitter voltage increases by 60mV. In the case of high-level injection, this increase is 120mV because of an additional factor 0.5 in the exponent of (5.17).

The temperature behavior

The temperature behavior of a junction voltage is easy to derive. This derivation is here more extensive because in a later paragraph the forward-biased junction is used as the core of a bandgap reference, and for this application the temperature behavior has to be well known. Equation (5.17) is changed into an equation with temperature-dependent variables:

$$I_C(T) = I_S(T) \exp\left[\frac{V_{BE}(T)q}{kT}\right]. \tag{5.20}$$

Then the temperature dependency is substituted for each variable. The temperature behavior of the collector current is determined by the bias current. It is assumed that the most convenient bias currents to realize are the currents with a temperature behavior equal to:

$$I_{\text{bias}} = I_{\text{bias}}(T_0)\left(\frac{T}{T_0}\right)^{\theta} \tag{5.21}$$

with T_0 a nominal temperature, $I_{\text{bias}}(T_0)$ the current at T_0 and θ the order of the temperature dependency (mostly θ is 0 or 1). The nominal temperature is the temperature from which later on the Taylor expansion is derived. The sources implementing this expression are treated in a later section.

For the temperature behavior of the saturation current, the following derivation suffices:

$$I_S(T) = \frac{qA{n_i}^2(T)\overline{D}(T)}{N_B} \tag{5.22}$$

in which A is the area of the junction, ${n_i}^2$ the intrinsic carrier concentration, $\overline{D}$ the mean minority-diffusion constant and N_B the Gummel number of the base region. For the intrinsic carrier concentration holds:

$$ {n_i}^2(T) = CT^3 \exp\left[-\frac{E_g(T)}{kT}\right] \tag{5.23}$$

with C a constant and $E_g(T)$ the bandgap energy as a function of the absolute temperature. The temperature dependency of $\overline{D}$ is found by using Einstein's relation. $\overline{D}(T)$ equals:

$$\overline{D}(T) = \frac{kT}{q}\overline{\mu}(T) \tag{5.24}$$

with $\overline{\mu}$ the mean mobility of the minority carriers in the base region. The temperature dependency of $\overline{\mu}$ can be defined as:

$$\overline{\mu}(T) = BT^{-n} \tag{5.25}$$

with B a constant and n the order of the temperature dependency. Putting all these equations together, the temperature dependency of the saturation current is given by:

$$I_S(T) = C'\left(\frac{T}{T_0}\right)^{\eta} \exp\left[\frac{-E_g(T)}{kT}\right] \tag{5.26}$$

with

$$\begin{aligned} \eta &= 4 - n, && (5.27)\\ C' &= \frac{T_0^{\eta} \cdot A \cdot B \cdot C \cdot k}{N_B}. && (5.28)\end{aligned}$$

The combination of (5.20), (5.26) and (5.21) yields the temperature dependency of the base-emitter voltage:

$$V_{BE}(T) = \frac{E_g(T)}{q} - (\eta - \theta)\frac{kT}{q}\ln\left(\frac{T}{T_0}\right) + \frac{kT}{q}\ln\left(\frac{I_{C0}}{C'{T_0}^{\theta}}\right) \tag{5.29}$$

with I_{C0} the collector bias current at the nominal temperature T_0. Two parameters, C' and I_{C0}, are eliminated when the equation for the base-emitter voltage is rewritten as:

$$V_{BE}(T) = V_{BE}(T) + \frac{T}{T_0}[V_{BE}(T_0) - V_{BE}(T_0)] \tag{5.30}$$

with $V_{BE}(T_0)$ the base-emitter voltage at the nominal temperature. For the temperature dependency of the base-emitter voltage a very convenient expression is found:

$$V_{BE}(T) = \frac{E_g(T)}{q} - \frac{T}{T_0}\left[\frac{E_g(T_0)}{q} - V_{BE}(T_0)\right] - (\eta - \theta)\frac{kT}{q}\ln\left(\frac{T}{T_0}\right). \tag{5.31}$$

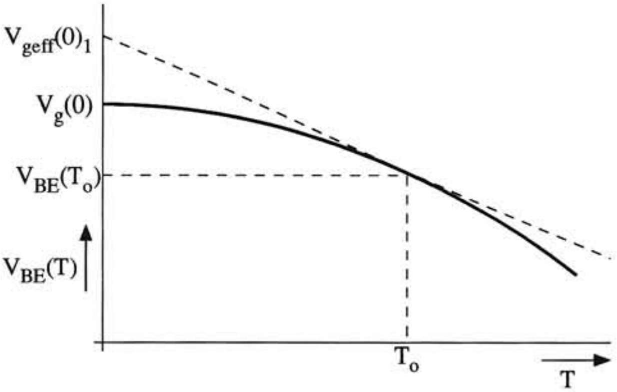


Figure 5.12: The base-emitter voltage as a function of the temperature

This function is plotted in figure 5.12. At 0K the base-emitter voltage equals the bandgap voltage at 0K, $V_g(0)$, and for increasing temperature the base-emitter voltage decreases slowly.

The Taylor polynomial is examined To find the first-order temperature behavior of the base-emitter voltage the Taylor polynomial is examined. The first-order Taylor polynomial around T_0 is given by:

$$V_{BE}(T)_1 = V_{BE}(T_0) - \left[V_g(0)_1 + \frac{kT_0}{q}(\eta - \theta) - V_{BE}(T_0)\right]\frac{T - T_0}{T_0} \qquad (5.32)$$

with $V_g(0)_1$ the bandgap voltage at 0K derived from a first-order Taylor polynomial of $V_g(T)$ near T_0. The temperature dependency of the base-emitter voltage is always negative and depends on the value of the base-emitter voltage. Further, the first-order approximation of the base-emitter voltage *always* intersects the y axis (T=0K) at:

$$V_{\text{geff}}(0)_1 = V_g(0)_1 + \frac{kT_0}{q}(\eta - \theta). \qquad (5.33)$$

Example:
The temperature behavior of several base-emitter voltages is calculated in this example. For the constants holds: the bandgap voltage $V_g(0)_1$ equals 1.2V (silicon), $T_0 = 300$K and the temperature dependency of the saturation current is of the third order. In the following cases the temperature dependency is calculated:

I) $V_{BE}(T_0) = 600$mV and the transistor is biased with a constant current;

II) $V_{BE}(T_0) = 600$mV and the transistor is biased with a current which is proportional to the absolute temperature, PTAT;

III) $V_{BE}(T_0) = 800$mV and the transistor is biased with a constant current.

Results:

I) The transistor is biased with a constant current and the temperature behavior of the saturation current is of the third order, thus $\theta = 0$ and $\eta = 3$. The first-order temperature behavior equals:

$$\frac{dV_{BE}}{dT} \mid_{T=T_0} = -2.259\text{mV/K}. \tag{5.34}$$

II) Now the transistor is biased with a PTAT current, i.e. $\theta = 1$, the other variables remain the same. The first-order temperature coefficient equals:

$$\frac{dV_{BE}}{dT} \mid_{T=T_0} = -2.173\text{mV/K}. \tag{5.35}$$

III) In this case only the base-emitter voltage is changed with respect to I:

$$\frac{dV_{BE}}{dT} \mid_{T=T_0} = -1.592\text{mV/K}. \tag{5.36}$$

From the above it may be clear that the temperature behavior of the base-emitter voltage is dependent on the biasing conditions of the transistor. For higher base-emitter voltages the first-order temperature coefficient reduces. This can be seen from figure 5.12. The slope of the tangent at $V_{BE}(T_0)$ decreases for higher base-emitter voltages as the intersection point of the tangent with the y axis remains at $V_{\text{geff}}(0)_1$, which is independent of $V_{BE}(T_0)$.

When the base-emitter junction is used as an accurate voltage reference, a high degree of accuracy can be obtained. The value of the base-emitter voltage at a certain temperature is given by equation (5.18) and a simplified expression of the temperature behavior is given by equation (5.32).

The accuracy of the resulting voltage is determined by the accuracy of the saturation current I_S and the collector bias current I_C. The accuracy of the saturation current is given by the accuracy of the emitter area. The larger the emitter area is, the higher the degree of accuracy, because of the decreasing relative influence of stochastic errors in diffusion and mask lithography. The accuracy of the collector bias current is determined by the accuracy of the current itself and the way in which the transistor is biased. When, for instance, a very accurate current source is used for biasing the collector current, and this source is also used to supply the base-current, the resulting accuracy may be poor. A proper way of biasing is depicted in Figure 5.13. The biasing is correctly done with the aid of a nullor. The nullor is an ideal circuit element. It controls its output current and voltage in such a way that the input voltage and current of the nullor become zero. In the circuit shown in figure 5.13, the nullor forces, by means of negative feedback, the

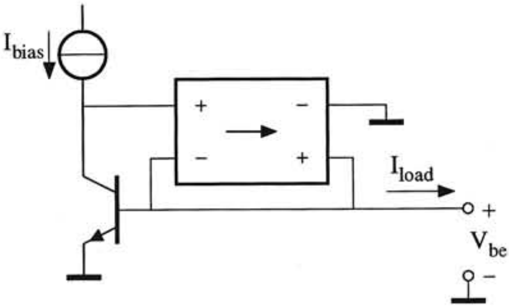


Figure 5.13: A proper way of biasing to get an accurate relation between V_{BE} and I_C

base-collector voltage to be zero. Thus, the forward Early effect can be neglected. Further, because of the zero input current of the nullor, the current I_{bias} from the current source flows completely through the collector.

The load current is supplied by the nullor also. A load current cannot influence the collector current. Thus, the base-emitter voltage is buffered, ideally, by the nullor.

A realistic circuit uses an approximation for the nullor. The simplest is just a wire as depicted in figure 5.14. In this configuration, the transistor is connected

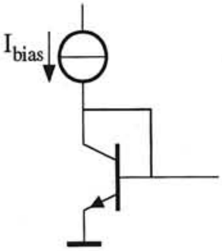


Figure 5.14: The nullor is implemented by a simple wire

as a diode and the base and load current are supplied by the collector bias source. This results in a difference between the actual collector current and the current supplied by the current source (the intended collector bias current).

As the transistor is diode connected, the output impedance r_{out} equals approximately:

$$r_{\text{out}} = \frac{1}{g_m} \tag{5.37}$$

where g_m is the transconductance of the transistor. When this impedance is too high, the bias current needs to be enlarged. When this is not possible, the nullor has to be implemented by amplifying stages, instead of the simple wire.

The noise performance

The noise performance of the base-emitter junction reference is found by transforming all the noise sources to the output. The noise performance is dominated by the thermal noise of the base resistance r_b and the collector shot noise. The power-density spectrum of the noise voltage is approximately:

$$S_{V_{BE}} = 4kT(r_b + \frac{1}{2g_m}). \tag{5.38}$$

Reduction of the noise is possible by choosing a transistor with a lower base resistance (this can be done by taking several transistors in parallel) or by choosing a higher collector current. Which noise source is dominant depends on the specific circuit. In low-power circuits, mostly the collector shot noise is dominant and the noise of the base resistance is negligible.

For instance, a typical value for the base resistance of a minimal sized transistor is 500Ω. When the transistor is biased at 1μA, the equivalent noise resistor at the output, representing the collector shot noise, equals 13kΩ. Thus the noise of the base resistance is negligible.

When low-noise voltage references are needed, the use of base-emitter junctions is the correct choice. This is easily seen when the noise of the base-emitter voltage reference is compared with the noise of a voltage source that is made with a resistor and a current source.

Example:
A reference of 600mV is made with a large transistor (for a high absolute accuracy). Assume 100μA is available for the biasing of the transistor and its base resistance is 150Ω. The total equivalent noise resistor equals 280Ω. This voltage can also be realized by a current flowing through a resistor. When the same bias current is used the required resistor equals:

$$R = \frac{U}{I} = \frac{0.6V}{100\mu A} = 6000\Omega. \tag{5.39}$$

Of course, the equivalent noise resistor is equal to this value. The noise power of the voltage reference made with the base-emitter junction is more than a factor 20 lower than the one made with the resistor.

The one-junction voltage reference can be very well used as a temperature sensor because of the good characterized temperature behavior. For instance, for a temperature dependency of approximately 2mV/K, a change in ambient temperature of 50K results in a change in the output voltage of 100mV. A voltage reference made by the difference of two junction voltages is discussed in the next section. The resulting expression for the temperature behavior is very accurate and simple.

5.3.4 The PTAT voltage source

The PTAT voltage source is a source with an output voltage which is proportional to the absolute temperature, or PTAT for short. Because of the proportionality to the absolute temperature, this voltage source is very well suited for use as a temperature sensor. The basis of a PTAT voltage source is the fact that the difference between two junction voltages is a PTAT voltage. The first section discusses a source that uses two transistors, each for one junction voltage. The following section treats a source where the two junctions in one transistor are used for the two required junction voltages, i.e. a saturating transistor.

Made with two transistors

The principle of a PTAT voltage using the difference between two base-emitter voltages is shown in figure 5.15. For the difference between two base-emitter voltages holds:

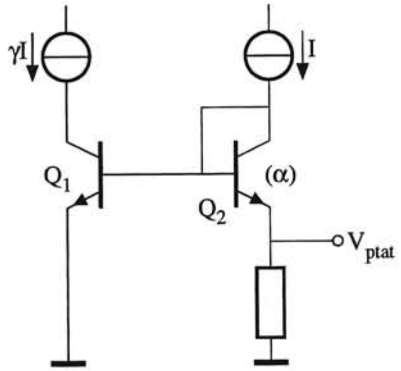


Figure 5.15: A PTAT voltage source using the difference between two base-emitter voltages

$$\Delta V_{BE} = V_{BE1} - V_{BE2} = \frac{kT}{q} \ln \left[\frac{I_{C1}(T)}{I_{C2}(T)} \cdot \frac{I_{S2}(T)}{I_{S1}(T)} \right]. \qquad (5.40)$$

When the two collector currents have the same temperature behavior and the temperature behavior of the two saturation currents are equal, the expression simplifies to:

$$\Delta V_{BE} = \frac{kT}{q} \ln \left(\frac{I_{C1}}{I_{C2}} \cdot \frac{I_{S2}}{I_{S1}} \right) = \frac{kT}{q} \ln(\gamma\alpha) \qquad (5.41)$$

with γ the ratio of the two collector currents and α the ratio of the two saturation currents. In figure 5.15, the collector currents of the two transistors are forced to have a ratio of γ. As the saturation current is proportional to the emitter area, α equals the ratio of the two emitter areas. From equation (5.41) follows that the difference between the two base-emitter voltages is PTAT.

Example: Assume the following holds for the collector and saturation currents:

- $I_{C1} : I_{C2} = 10 : 1$,
- $I_{S1} = I_{S2}$.

Then the PTAT voltage equals:

$$\Delta V_{BE} = \frac{kT}{q} \ln(10) = 199 \cdot \mu\text{V/K} \cdot T = 59.6\text{mV @ 300K}. \qquad (5.42)$$

Thus for each degree Kelvin the temperature changes, ΔV_{BE} changes 199μV.

The output impedance, r_{out}, of this source equals:

$$r_{\text{out}} = \frac{1}{g_{m1}} + \frac{1}{g_{m2}} = \frac{kT}{qI}\left(\frac{1+\gamma}{\gamma}\right) \qquad (5.43)$$

with g_{m1} and g_{m2} the transconductance of Q_1 and Q_2, respectively.

The power-density spectrum, S_u, of the PTAT voltage is given by the sum of the noise from the two base-emitter voltages and equals:

$$S_u = 4kT\left(r_{b1} + r_{b2} + \frac{0.5}{g_{m1}} + \frac{0.5}{g_{m2}}\right). \qquad (5.44)$$

Using a saturated transistor

A very simple PTAT voltage source is the one that uses one saturated transistor. In this source, the two junction voltages are the base-emitter and the base-collector voltage. The difference between those two junction voltages is the collector-emitter voltage. The circuit is shown in figure 5.16. The transistor is saturated by forcing a base current into the base for which holds:

$$I_B \geq \frac{I_C}{\beta_f} \qquad (5.45)$$

with β_f the current-gain factor in the normal forward mode. The base-collector junction is also biased in the forward region. Thus, both junctions are conducting junctions and the collector-emitter voltage is determined by the difference between two junction voltages. This is depicted in figure 5.17. The collector-emitter voltage is given by:

$$V_{CE} = \frac{kT}{q} \ln\left(\frac{1 + 1/\beta_r + I_C/I_B\beta_r}{1 - I_C/I_B\beta_f}\right) \qquad (5.46)$$

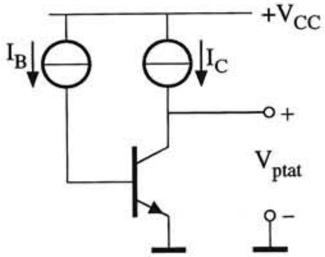


Figure 5.16: A PTAT voltage source using a saturated transistor

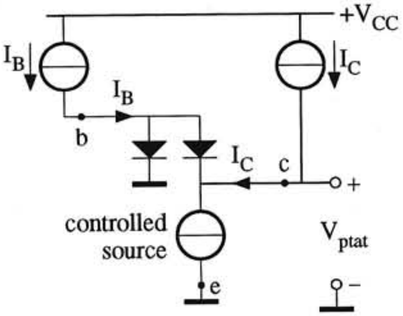


Figure 5.17: The collector-emitter voltage of a saturated transistor

with β_r and β_f the reverse and forward current-gain factor in the normal regions, respectively. Both the collector current and the base current are forced into the transistor. This results in a "forced current-gain factor" β_{sat} of the saturated transistor and is defined by:

$$\beta_{\text{sat}} = \frac{I_C}{I_B}. \tag{5.47}$$

Substituting β_{sat} into the expression for the collector-emitter voltage, yields:

$$V_{CE} = \frac{kT}{q} \ln \left(\frac{1 + 1/\beta_r + \beta_{\text{sat}}/\beta_r}{1 - \beta_{\text{sat}}/\beta_f} \right). \tag{5.48}$$

Thus the collector-emitter voltage of a saturated transistor is PTAT. It is assumed that the three current-gain factors are temperature independent.

Example:
When for the three current-gain factors hold: $\beta_r = 3$, $\beta_f = 100$, $\beta_{\text{sat}} = 20$, the collector-emitter voltage equals:

$$V_{CE,\text{sat}} = 199\mu\text{V/K} \cdot T = 59.6\text{mV @ 300K}. \tag{5.49}$$

For higher output voltages, several saturating transistors may be stacked.

The small-signal output impedance of this source is given by the derivative of the collector-emitter voltage with respect to I_C, resulting in:

$$r_{\text{out}} = \frac{kT}{qI_C} \left(\frac{\beta_f}{\beta_f - \beta_{\text{sat}}} \right) \frac{1}{1 + (\beta_r + 1)/\beta_{\text{sat}}}. \tag{5.50}$$

For $\beta_{\text{sat}} \ll \beta_f$ and $\beta_r \ll \beta_{\text{sat}}$ this reduces to:

$$r_{\text{out}} = \frac{kT}{qI_C}. \tag{5.51}$$

The output impedance is similar to the impedance of a diode connected transistor and thus can be relatively low. For instance, for a collector current of 1mA the output impedance equals 25Ω.

The power-density spectrum, S_u, of the output noise voltage is again determined by the noise generated by the two junctions. S_u is given by:

$$S_u = 4kT \left(0.5 \frac{kT}{qI_{BE}} + 0.5 \frac{kT}{qI_{BC}} \right) \tag{5.52}$$

with I_{BE} the current flowing through the base-emitter junction and I_{BC} the current flowing through the base-collector junction.

5.3.5 The bandgap reference

When temperature-independent voltages are needed, the combination of a junction at avalanche breakdown and a forward-biased junction may suffice. However, the minimally required supply voltage is about 7V. In the growing area of low-voltage electronics, with supply voltages down to 1V, this kind of references is not feasible. Other types of references have to be used. A circuit that can still work at those low supply voltages is the bandgap reference.

A bandgap reference is a voltage source of which the output voltage is related to the bandgap voltage at 0K. Because this voltage is a constant, the output voltage of a bandgap reference is ideally temperature independent.

Equation (5.31) was found for the temperature behavior of the base-emitter voltage. This equation can be represented by a Taylor series around a nominal temperature T_0 as:

$$V_{BE}(T) = \alpha_0 + \alpha_1(T - T_0) + \alpha_2(T - T_0)^2 + \cdots \tag{5.53}$$

The objective of bandgap reference design is to cancel the temperature coefficients of this base-emitter voltage. This is possible by taking an appropriate linear combination of base-emitter voltages:

$$\sum_{i=1}^{n} a_i V_{BE_i}(T) = \sum_{i=1}^{n} a_i \alpha_{0_i} + \sum_{i=1}^{n} a_i \alpha_{1_i}(T - T_0) + \sum_{i=1}^{n} a_i \alpha_{2_i}(T - T_0)^2 + \cdots \tag{5.54}$$

To obtain a temperature-independent reference voltage, V_{ref}, the first term on the right-hand side has to be equal to V_{ref} and the other terms on the right-hand side need to be zero. In this case the reference voltage is totally temperature independent. However, in most cases this is an overkill. Reasonable results can be obtained when only the first-order behavior is canceled, because this is by far the largest disturbing factor.

The linear combination of base-emitter voltages, an implicit compensation

In this section the linear combination of base-emitter voltages, in order to obtain a first-order compensated bandgap reference, is discussed. Because second and higher-order terms are not considered, only the first two terms of the Taylor series are used. For these two terms expression (5.32) was found. This equation can be rewritten to the convenient expression:

$$V_{BE}(T)_1 = V_{\text{geff}}(0)_1 - [V_{\text{geff}}(0)_1 - V_{BE}(T_0)]\left(\frac{T}{T_0}\right) \tag{5.55}$$

with $V_{\text{geff}}(0)_1 = V_g(0)_1 + \frac{kT_0}{q}(\eta - \theta)$ the effective bandgap voltage. This expression clearly shows the relation between the base-emitter voltage and the bandgap voltage.

Because only the constant term and the first-order behavior of the output voltage needs to be set, a linear combination of two base-emitter voltages is sufficient. The block diagram is depicted in figure 5.18. The linear combination is given by:

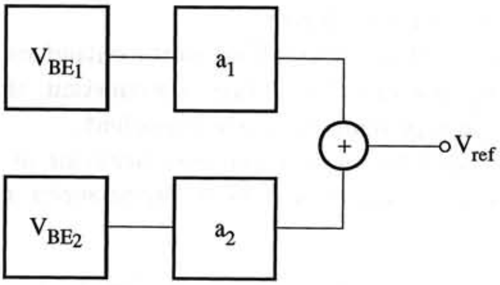


Figure 5.18: A linear combination of two base-emitter voltages

$$V_{\text{ref}} = (a_1+a_2)V_{\text{geff}}(0)_1 - [(a_1+a_2)V_{\text{geff}}(0)_1 - a_1 V_{BE_1}(T_0) - a_2 V_{BE_2}(T_0).]\left(\frac{T}{T_0}\right) \quad (5.56)$$

This equation may make the name "bandgap reference" clear: the reference voltage is directly related to the bandgap voltage and is given by:

$$V_{\text{ref}} = (a_1 + a_2)V_{\text{geff}}(0)_1. \quad (5.57)$$

The reference voltage can be set to all kinds of values by choosing the sum of the two scaling factors. Substitution of this expression in the first-order part of equation (5.56) results in a constraint for the first-order compensation:

$$a_1 V_{BE_1}(T_0) + a_2 V_{BE_2}(T_0) = V_{\text{ref}}. \quad (5.58)$$

From equations (5.57) and (5.58), the two scaling factors can be found to be:

$$a_1 = +\frac{V_{\text{ref}}}{V_{\text{geff}}(0)_1} \cdot \frac{V_{\text{geff}}(0)_1 - V_{BE2}(T_0)}{V_{BE1}(T_0) - V_{BE2}(T_0)}, \quad (5.59)$$

$$a_2 = -\frac{V_{\text{ref}}}{V_{\text{geff}}(0)_1} \cdot \frac{V_{\text{geff}}(0)_1 - V_{BE1}(T_0)}{V_{BE1}(T_0) - V_{BE2}(T_0)}. \quad (5.60)$$

These equations show that the two scaling factors have an opposite sign. This is because the first-order temperature coefficient of a base-emitter voltage is always negative. To obtain first-order compensation, the *difference* between the two base-emitter voltages has to be taken. Further, these two expressions show that the two base-emitter voltages have to be *different*. This is because equal base-emitter voltages have the same temperature dependency. Performing a first-order temperature compensation with equal V_{BE}s, and thus having the same temperature dependency, would result in a constant term equal to zero which is impractical.

The principle of the first-order temperature compensation with a linear combination of two $V_{BE}s$, is depicted in figure 5.19. In the figure, the base-emitter voltages do have a different value at T_0 and thereby a different first-order behavior. The remaining temperature behavior is of the second and higher order.

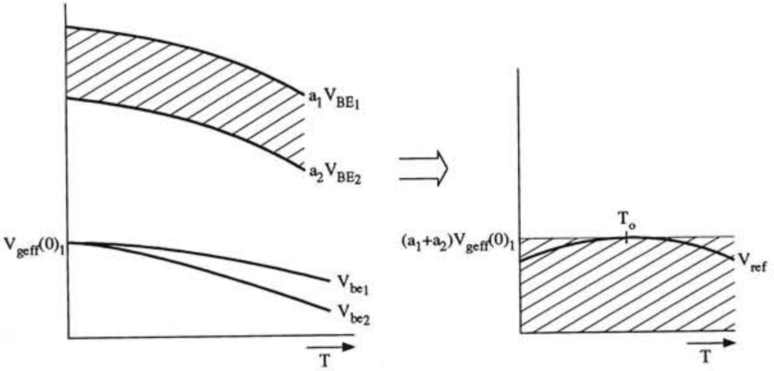


Figure 5.19: The principle of the compensation with a linear combination

The noise behavior

In section 5.3.3, the noise behavior of a single-diode reference was discussed. In the previous section it was shown that the bandgap reference is just a weighted summation of a number of those diode references (base-emitter voltages) . The power-density spectrum, S_v, of the noise voltage at the output of a bandgap reference is therefore given by:

$$S_v = 4kT \left[a_1^2 \left(r_{b1} + \frac{1}{2g_{m1}} \right) + a_2^2 \left(r_{b2} + \frac{1}{2g_{m2}} \right) \right]. \tag{5.61}$$

The noise is scaled by the same factor as the corresponding base-emitter voltages. Again, two types of noise sources are involved. The thermal noise of the base-resistances and the shot noise of the collector current. The former one can be lowered by using larger transistors with a lower base resistance. The latter can be decreased by using more current (g_m increases). For optimal use of the current with respect to noise, the influence of the base resistances has to be made negligible. This can easily be done for bias currents up to several 100μAs In this case the total noise (the implementations of the scaling factors a_1 and a_2 are assumed to be noiseless) is only from collector shot noise. When a noise minimization is done with the constraints of a first-order temperature-compensated bandgap reference and a limitation on the current consumption, the following holds:

$$\ln\left(x\frac{A_2}{A_1}\right) = -2\frac{1+x}{1-x} \tag{5.62}$$

with A_1 and A_2 the two emitter areas and x the ratio of I_{C1} and I_{C2}, the two collector currents. This expression states that:

For a given ratio of the two emitter areas, an optimal ratio of the two collector bias currents exists, for which the noise of the first-order compensated bandgap reference is minimal. This ratio is independent of the total current consumption $I_{C1} + I_{C2}$.

A design example

In the previous sections, the theory of the bandgap reference was discussed using ideal $V_{BE} - I_C$ relations. In this section, a design example of a bandgap reference is presented using the Gummel and Poon model for $I_C = f(V_{BE})$. The $V_{BE} - I_C$ relation becomes more realistic, and the design aims at a minimization of the number of non-idealities that have to be taken into account. According to the Gummel and Poon model, the relation between the base-emitter voltage and the collector current for a normal biased transistor is:

$$I_C(T) = \left(1 - \frac{V_{BC}}{V_{AF}} - \frac{V_{BE}}{V_{AR}}\right) \cdot \frac{I_{BE1}}{\frac{1}{2}\left[1 + (1 + 4I_{BE1}/I_{KF})^{NK}\right]} \tag{5.63}$$

with I_{BE1} defined as:

$$I_{BE1} = I_S(T) \cdot \exp\left(\frac{V_{BE}}{kT/q}\right) \tag{5.64}$$

and I_S according to:

$$I_S(T) = I_S(T_0) \cdot \exp\left[\left(\frac{T}{T_0} - 1\right)\frac{E_G}{kT}\right] \cdot \left[\frac{T}{T_0}\right]^{X_{TI}} . \tag{5.65}$$

When the transistor is biased far below high-level injection ($I_{BE1} \ll I_{KF}$) and the base-collector voltage is kept zero ($V_{BC} = 0$), (5.63) reduces to:

$$I_C(T) = \left(1 - \frac{V_{BE}}{V_{AR}}\right) \cdot I_S(T)\exp\left(\frac{V_{BE}}{kT/q}\right). \tag{5.66}$$

This is the same equation as (5.20) except for V_{AR} and some differences in the expression for $I_S(T)$. In the Gummel and Poon model, the parameter for the temperature behavior of I_S is X_{TI} instead of η. Further, the model uses a first-order temperature model for the bandgap energy.

V_{AR} is the reverse Early voltage. This parameter is used for describing the base-width modulation at the base-emitter junction. This Early voltage can be low, i.e. several volts. The errors in the base-emitter voltages, due to the V_{AR}, can be transformed to the output of the bandgap reference. The resulting error in the reference voltage equals:

$$V_{\text{error}} = \frac{kT}{q}\frac{V_{\text{ref}}}{V_{AR}}. \tag{5.67}$$

This error can be accounted for by adding to, for instance V_{BE1}, a term

$$V_{\text{additional}} = \frac{V_{\text{error}}}{a_1} \tag{5.68}$$

and by solving again the set of equations resulting from the linear combination.

From the foregoing, it appears that four parameters have to be known accurately: the key parameters. The other parameters have to be kept either as large or as small as possible. The four key parameters are:

- E_G the bandgap energy,
- I_S the saturation current,
- X_{TI} the order of temperature behavior of the saturation current,
- V_{AR} the reverse Early effect.

Now the blocks of the bandgap references need to be designed such that the simplifications made are valid. To obtain a base-emitter voltage according to relation (5.66) the topology of figure 5.20 can be used (c.f. figure 5.13). The nullor is implemented by a single MOS differential pair. The MOS transistor is favorable here because of the absence of a gate current so that the bias current I_C flows completely through the collector lead.

This cell is used as the core of a first-order compensated bandgap reference with an output voltage equal to $V_{\text{geff}}(0)_1$. According to (5.57), the sum of the two scaling factors is given by:

$$a_1 + a_2 = 1. \tag{5.69}$$

Because the sum of the two scaling factors equals one, equation (5.58) can be rewritten as:

$$\begin{aligned} V_{\text{ref}} &= (a_1 + a_2)V_{BE1}(T_0) + a_2[V_{BE2}(T_0) - V_{BE1}(T_0)] \\ &= V_{BE1}(T_0) + a_2[V_{BE2}(T_0) - V_{BE1}(T_0)]. \end{aligned} \tag{5.70}$$

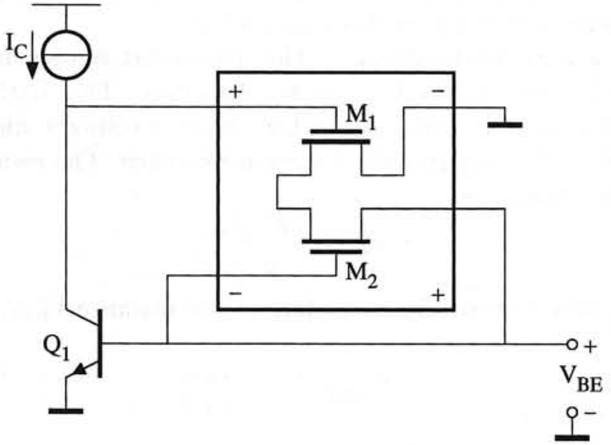


Figure 5.20: Generation of a base-emitter voltage

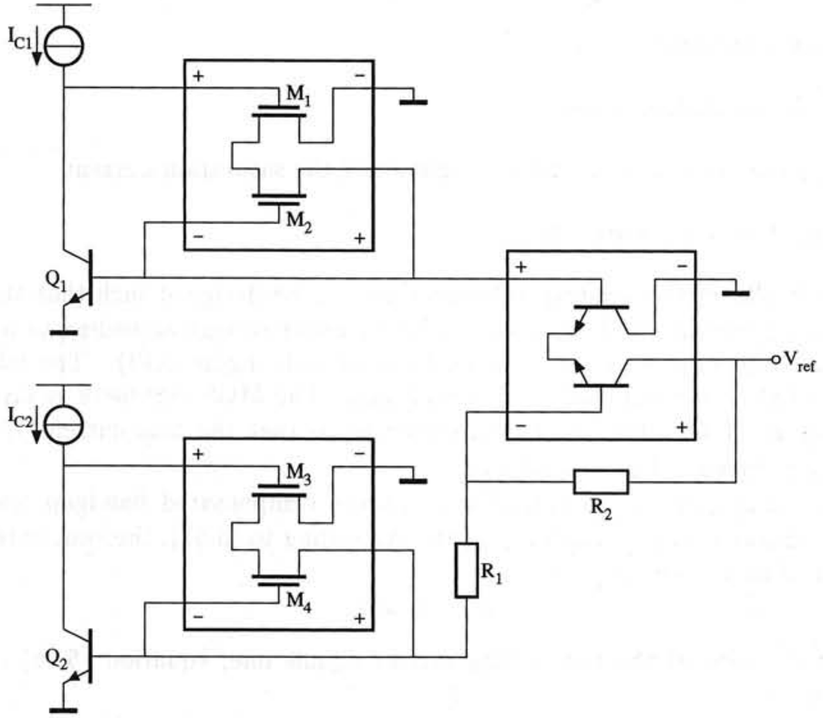


Figure 5.21: The topology of the bandgap reference circuit

The bandgap reference can thus be made by the topology as depicted in figure 5.21. The voltage scaler a_2 is realized by a bipolar differential pair, and resistors R_1 and R_2. The scaling factor is given by

$$a_2 = 1 + \frac{R_2}{R_1} \tag{5.71}$$

Here, a bipolar differential pair is used because of the low input-offset voltage. This voltage is directly in series with the reference voltage and thus needs to be as low as possible.

The two collector bias currents can be derived from a PTAT voltage source via a resistor. The resistor needs to be accurate because its absolute value is important. The ratio of the two currents is given by the equation for noise minimization.

There are a few important design aspects, with respect to integration, to pay attention to:

- Use large reference transistors for optimal matching. Take care that the effective emitter areas are equal (emitter crowding).
- Use relatively large-sized resistors for accurate scaling factors. The scaling factor depends only on the matching of the resistors.
- Place components to be matched close to each other.

Unfortunately, the total error in the output voltage of the bandgap reference may still be such that trimming is needed to obtain a temperature-independent voltage again. For this purpose R_1 or R_2 needs to be adjustable.

The conventional way, an explicit compensation

In the previous section, the temperature behavior of a base-emitter voltage was compensated implicitly. Two scaled base-emitter voltages were added to set the output voltage and, at the same time, compensate the first-order temperature behavior. Both voltages have a constant and a first-order term. In this section, the temperature behavior is compensated explicitly. The compensation of the first-order temperature behavior of a base-emitter voltage is done by a separate PTAT voltage which is added to the base-emitter voltage. The principle of this method is depicted in figure 5.22. From this figure, it can be seen what has to be done to obtain a temperature-independent voltage equal to $V_{\mathrm{geff}}(0)_1$. A voltage, represented by the shaded area, has to be added to a base-emitter voltage. This compensation voltage is zero at 0K and increases linearly for increasing temperature. Thus the compensation voltage needs to be PTAT. When the y axis is scaled by a factor a, an output voltage, V_{out}, equal to:

$$V_{\mathrm{out}} = aV_{\mathrm{geff}}(0)_1 \tag{5.72}$$

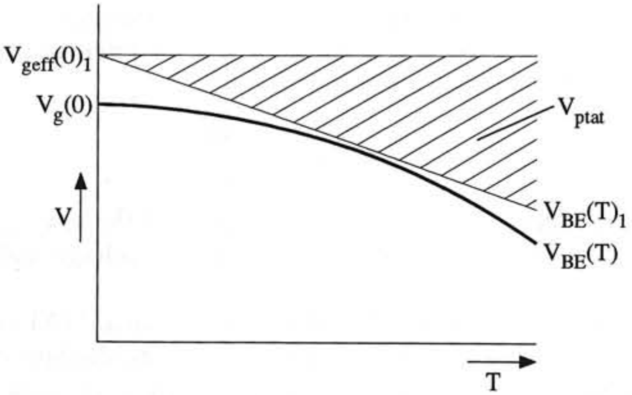


Figure 5.22: Explicit compensation of the temperature behavior of a base-emitter voltage

is obtained. In this case, the temperature behavior of a fraction a of a base-emitter voltage has to be compensated.

Assume a reference voltage equal to $V_{\text{geff}}(0)_1$ is required and transistor II from section 5.3.3 is used:

- $V_{BE}(T_0) = 600\text{mV}$,
- $\alpha_{1_{VBE}} = -2.173\text{mV/K}$ (transistor is biased with a PTAT current).

To compensate α_1, a PTAT voltage with a first-order temperature dependency of 2.173mV/K is required. In section 5.3.4, it was calculated that for a PTAT voltage source with emitter scaling one and current scaling ten holds:

- $V_{\text{PTAT}}(T_0) = 59.6\text{mV}$,
- $\alpha_{1,\text{PTAT}} = 199\mu\text{V/K}$.

To attain total compensation, the PTAT voltage needs to be amplified by an factor A_V of :

$$A_V = \frac{-\alpha_{1,VBE}}{\alpha_{1,\text{PTAT}}} = \frac{2.173mV/K}{198\mu V/K} = 10.92. \tag{5.73}$$

When the amplified PTAT voltage is added to the base-emitter voltage, the reference voltage equals at 300K:

$$V_{\text{out}} = V_{BE} + A_V V_{\text{PTAT}} = 0.6\text{V} + 10.92 \cdot 59.6\text{mV} = 1.25\text{V} \tag{5.74}$$

and equals to:

$$V_{\text{out}} = V_{\text{geff}}(0)_1 = V_g(0)_1 + \frac{kT_0}{q}(X_{TI} - \theta) = 1.25\text{V}. \tag{5.75}$$

This voltage has a temperature coefficient of 0mV/K at the nominal temperature 300K. The block diagram of the bandgap reference is depicted in figure 5.23.

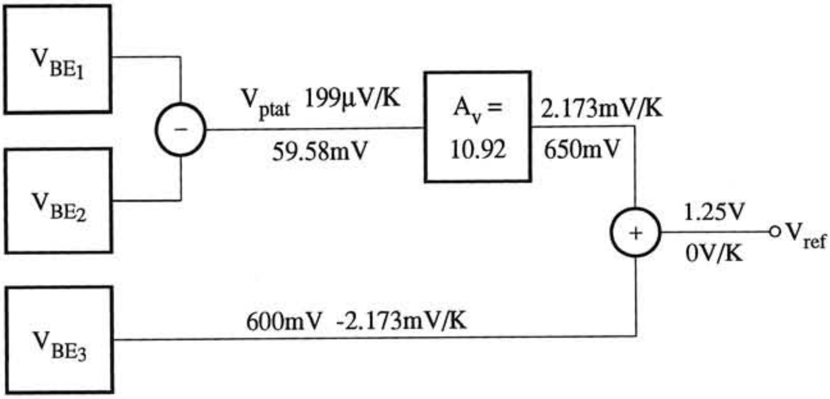


Figure 5.23: The block diagram of a bandgap reference with explicit compensation

Accuracy aspects

The design of accurate bandgap references relies on the compensation of one temperature-dependent voltage with another, thus, matching is evidently of prime importance. This can be seen from the expression describing the output voltage:

$$\begin{aligned} V_{\text{ref}} &= (a_1 + a_2)V_{\text{geff}}(0)_1 \\ &- \{a_1[V_{\text{geff}}(0)_1 - V_{BE1}(T_0)] + a_2[V_{\text{geff}}(0)_1 - V_{BE2}(T_0)]\}\frac{T}{T_0} \end{aligned} \quad (5.76)$$

The accuracy of several parameters is important:

- The scaling factors a_1 and a_2. These have to be implemented by ratios made of components with a good matching.
- $V_{BE1}(T_0)$ and $V_{BE2}(T_0)$. The transistors used should have matched emitter areas and the two collector bias currents should be derived with the help of matching.
- $V_g(0)$ and X_{TI}, via $V_{\text{geff}}(0)_1$. These are process parameters and their accuracy depends on process stability and, of course, on the accuracy of the parameter extraction.

Several parameters are involved in the design of the bandgap reference. All of them have more or less stochastic spread. Thus, one of the two scaling factors needs to be trimmed to account for this spread. Moreover, there is one more component that has a rather great influence up on the accuracy. This is the resistor by which the bias currents are related to a well-known voltage (most easy is a PTAT voltage). Its absolute value is important. Resistors on chip have an absolute accuracy of about 10 to 20%. The error caused by this uncertainty can easily by the major reason for trimming.

The design of bandgap references in MOS processes

There are two general ways of making a bandgap reference in a MOS process. Firstly, the parasitic substrate PNP can be used. See figure 5.24. The drawback of this method is the uncertainty in the collector current. For accurate bandgap reference design the relation between the base-emitter voltage and the collector current has to be used. In the case of a parasitic PNP transistor, the collector current is determined *indirectly* via the emitter current. The base current now influences the behavior of the base-emitter voltage.

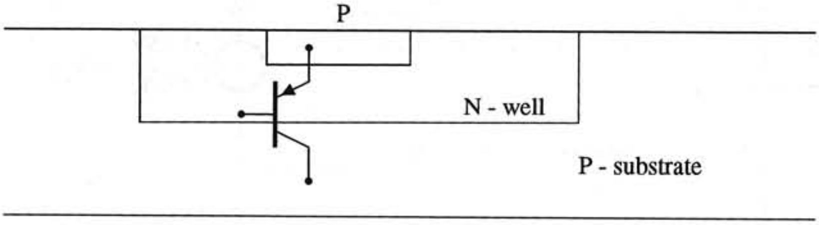


Figure 5.24: A parasitic PNP in a MOS process

Secondly, the bandgap reference can be made by using MOS transistors in weak inversion. In this case, the relation between the gate-source voltage and the drain current is exponential, as it is for bipolar transistors. The bandgap energy occurs in the same way in this relation as it does in the case of bipolar transistors. The problem is in the definition of the gate-source voltage. This is depicted in figure 5.25. The effective gate-source voltage is the voltage that is across the channel, V_{channel}. This effective voltage is related to the external voltage by the capacitive division of C_{ox} and C_{channel}. These two parameters directly introduce an uncertainty in the gate-source voltage. For accurate design, this ratio has to be well known.

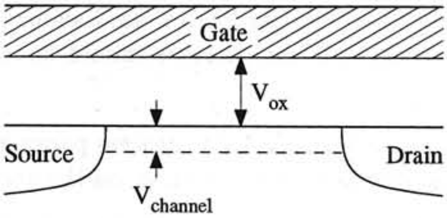


Figure 5.25: The gate-source voltage of a weak inversion MOS transistor

Therefore, currently, for *accurate* bandgap reference design, a bipolar transistor process is best suited.

The influence of stress in a chip

After the wafer with bandgap references returns from the ic foundry, the bandgap reference is trimmed to obtain the low temperature dependency. Subsequently, the wafer is sawed into separate chips. Each chip contains a bandgap reference. Next, these chips are mounted in a package with glue, and the total package is heated to dry the glue. The problem arises in this last step. Due to the different thermal behaviors of the glue and the silicon, during this heating step stresses arise in the chip and the behaviors of the devices on the chip change slightly. Now the bandgap reference is no longer optimally trimmed and a second trim procedure has to be performed. This stress seems to depend on the orientation of the molecule lattice with respect to the surface, e.g. a perpendicular orientation or under a certain angle. The former is less sensitive to stresses than the latter.

Some concluding remarks

In this section, the first-order compensated bandgap reference has been discussed. With this reference, temperature dependencies of about some tens of ppm/K over a range of 100K can be obtained. When lower dependencies are needed, higher orders of the temperature behavior also have to be compensated for. These bandgap references, mostly called second-order compensated or curvature-corrected references, can have temperature dependencies of only 1ppm/K over a range of 100K.

5.3.6 Conclusions on voltage sources

In the foregoing sections, several ways of implementing voltage constants were discussed. The simplest implementation derives a voltage from the supply voltage by means of a resistive divider. This source consumes a rather high bias current to obtain a low output impedance and a low noise level.

The source using a non-linear device in the divider has a better performance. The forward-biased junction has shown to be a very good candidate for this non-linear device. A low-noise behavior combined with a low output impedance is feasible.

A voltage source with an output voltage which is Proportional To the Absolute Temperature (PTAT) is obtained when the difference between two junction voltages is used. The junctions may be from two separate transistors or from one saturating transistor, i.e. the base-emitter junction and the base-collector junction.

Finally, the bandgap reference was treated. This source has an output voltage that is related to the bandgap voltage at 0K and has, consequently, a temperature coefficient of 0V/K, ideally. When constructed with a linear combination of two base-emitter voltages, there is an optimal bias-current ratio for the two base-emitter voltage generators with minimum noise level. For the implementation of the bandgap reference, the absolute value of at least one resistor is important, which makes trimming inevitable.

5.4 The current source

In electronics there are voltage references (i.e. the bandgap voltage), currents references, however, do not exist because of the absence of a magnetic monopole. The current references which are realized are all derived from a voltage reference with the aid of a (trans)conductance. This is depicted in figure 5.26. When the voltage source is assumed to be ideal, the power-density spectrum S_i of the noise current at the output equals:

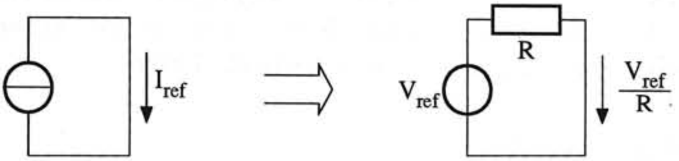


Figure 5.26: Derivation of a current source from a voltage source

$$S_i = \frac{4kT}{R} \tag{5.77}$$

and can ideally be zero when a source with an infinitely high voltage is used such that an infinitely high resistance can be used. But in practice, the noise is worse than given in equation (5.77). This is because a current flowing through a resistor induces 1/f noise due to the granular structure of the resistive material. Especially in high ohmic resistors is the structure of the material relatively rough, for instance, in poly-silicon resistors. In metal film and diffused resistors this effect is of less influence.

When both high output impedances and high output currents are needed, this type of current source gives problems because high reference voltages are required. It is better to realize a high output impedance with an active circuit.

5.4.1 An active current source using a transconductance

The basic configuration of an active current source is given in figure 5.27 in which a nullor is used. The current is again given by:

$$I_{\text{ref}} = \frac{V_{\text{ref}}}{R}. \tag{5.78}$$

The output impedance is enlarged by the negative feedback, and is actually infinite due to the nullor. Ideally, the output impedance is independent of the resistor R.

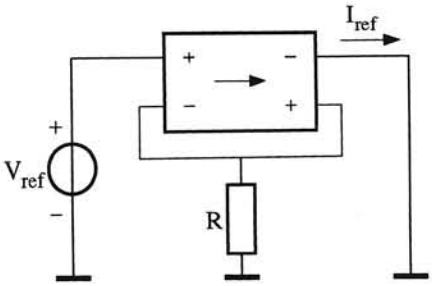


Figure 5.27: A current source realized by an active transconductane amplifier and a reference voltage

Using bipolar transistors

When the simplest implementation for the nullor is used, a CE stage, the series stage of figure 5.28 is obtained. The output current is given by (the base current is ignored):

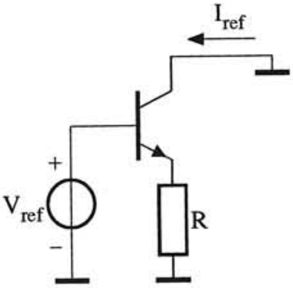


Figure 5.28: The nullor implemented by a single CE stage

$$I_{\text{out}} = \frac{V_{\text{ref}} - V_{BE}}{R}. \tag{5.79}$$

As the base-emitter voltage is temperature dependent, the output current of the source is temperature dependent also. Resistor R performs a feedback action, and through this a higher output impedance is obtained. To see what the increase of output impedance is as a function of the feedback (or loop gain), the output impedance is calculated. The output impedance equals:

$$r_{\text{out}} = r_o + R + R\frac{\beta_f r_o - R}{r_b + r_\pi + R}. \tag{5.80}$$

This function is plotted in figure 5.29. The output impedance is normalized to r_o and the feedback resistor to r_π. For values of R much larger than r_π plus r_b the expression for the output impedance reduces to:

$$r_{\text{out}} = (\beta_f + 1)r_o. \tag{5.81}$$

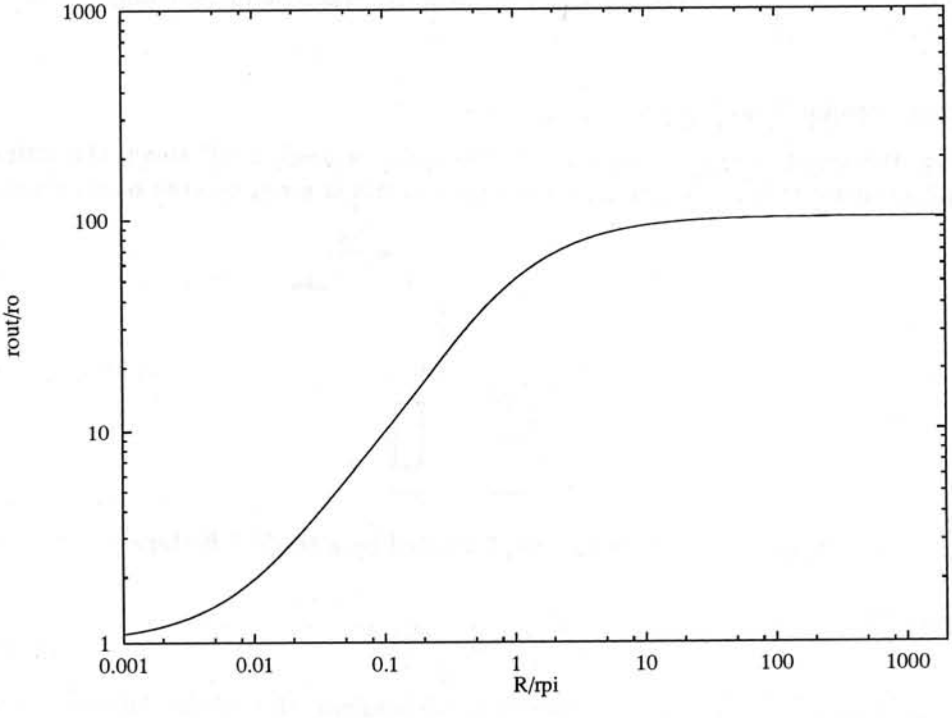


Figure 5.29: The normalized output impedance as a function of the normalized feedback resistor for β=100

Usually, this results in very high values for R. For 70% of the maximal output impedance, R needs to be approximately 2 times larger than r_π. Then, the voltage across the feedback resistor equals:

$$V_R = I \cdot R = I \cdot 2r_\pi = I \cdot 2\beta \frac{kT}{qI} = 2\beta \frac{kT}{q} \tag{5.82}$$

For $\beta = 100$ and at room temperature the voltage across the resistor is about $5V$. This 5V constitutes a limitation for this type of current source. Because, in the increasing area of low-voltage design, i.e. a supply voltage of 1V, this current source cannot be so implemented. Another limitation is given by the required resistance. In low-current applications, the output current can easily be in the order of nAs or μAs, which is demanding for feedback resistors in the order of MΩs and GΩs.

Increasing the output impedance of the source shown in figure 5.28, without the need for very high resistances, can be realized in two ways:

- Increase of the loop gain ($\beta + 1$). Output current *variations* due to an output voltage variations are more suppressed. A better approximation is made for the nullor.
- Increase of the output impedance of the active part (r_o). Now the output impedance without loop gain is already higher.

For the first option, the nullor is implemented, for instance, with a two-stage bipolar amplifier. The output impedance increases by a factor β. The second option is obtained when the CE stage is cascoded, see figure 5.30. Transistor Q' is the cascode transistor. It is a current buffer for the output of the CE stage,

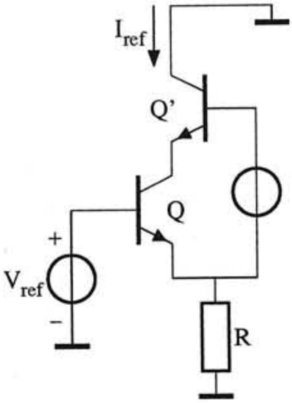


Figure 5.30: An active current source with a cascoded transistor

transistor Q. Without feedback, the output impedance of the buffered CE stage is approximately βr_o. When subsequently the feedback action is added to the current source, the output impedance can increase even further to approximately $\beta^2 r_o$. When a still higher output impedance is required, the loop gain has to be increased. An additional cascode transistor does not help.

Using FETs

The current source with one FET is depicted in figure 5.31. The nullor shown in figure 5.27 is now implemented with a CS stage. The output resistance equals:

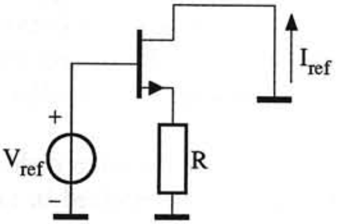


Figure 5.31: An active current source with a FET

$$r_{\text{out}} = R + (1 + g_m R) r_d \tag{5.83}$$

with r_d the small-signal output resistance of the FET and g_m its transconductance factor. In contrast with the output impedance for the bipolar source, which is limited to $(\beta_f + 1) r_o$, the output impedance of the FET source tends to infinity for a feedback resistor tending to infinity. But, again, high resistances are required. To increase the output impedance, without the need for high resistances, two options are possible:

- Make a better approximation for the nullor;
- Enlarge r_d by means of cascoding.

An example of the second option is depicted in figure 5.32. The output impedance of the cascoded CS stage without the feedback is now:

$$r_{d,\text{cascode}} = g_m r_d \cdot r_d = \mu r_d. \tag{5.84}$$

The output impedance is increased by a factor equal to the voltage-gain factor of the FET. For each additional cascode FET, the output impedance increases a factor μ, in contrast to the bipolar implementation where a second cascode transistor does not help.

The difference in the behavior of the current source made by the bipolar transistor and the FET is caused by the nature of the effect that causes the finite output impedance of these devices:

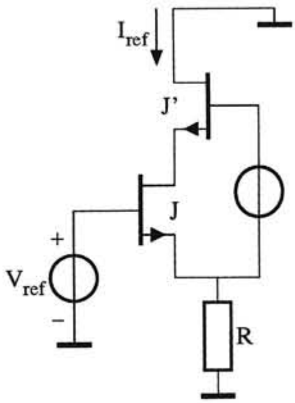


Figure 5.32: A FET current source with a cascoded FET

- Bipolar: r_o is caused by base-width modulation at the *base-collector* junction;
- FET: r_d is caused by the channel-length modulation, i.e. a *source-drain* effect.

In the case of the bipolar transistor, a part $1/\beta_f$ of the current (the base current) leaks away via the base terminal. This leakage current is not seen by the feedback. The FET does not have this "leakage". Theoretically an infinite impedance can be obtained with only cascoding.

Noise behavior

In this section, the influence of the amount of feedback on the noise behavior is discussed. The current source with all its noise sources is depicted in figure 5.33. The power-density spectrum S_i of the equivalent noise current at the output is given by:

$$S_i = \frac{2qI_C}{(1+\beta R/r_\pi)^2} + \frac{2qI_B}{(1+r_\pi/\beta R)^2} + \frac{4kT(R+r_b)}{(R+r_\pi/\beta)^2}. \tag{5.85}$$

The first term represents the effect of the collector shot noise. For very high feedback-resistor values, this term disappears completely. The second term is due to the base shot noise. For very high values for R this term reduces to $2qI_B$. The last term accounts for the noise of the feedback resistor and the base resistance. This term vanishes for high values of the feedback resistor. The function is plotted in figure 5.34. The noise-power density has been normalized to $2qI_C$ and the feedback resistor has been normalized to r_π.

To obtain the minimum noise-power density level, of $2qI_B$, a rather high value for the feedback resistor is required (c.f. the discussion about increasing the output

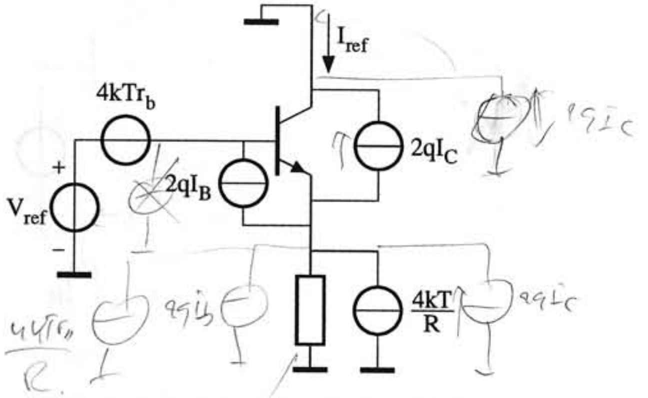


Figure 5.33: The noise sources in the active current source

R=0 → 2qIc out.
R=∞ ⇒ 2qIc cancelled itself

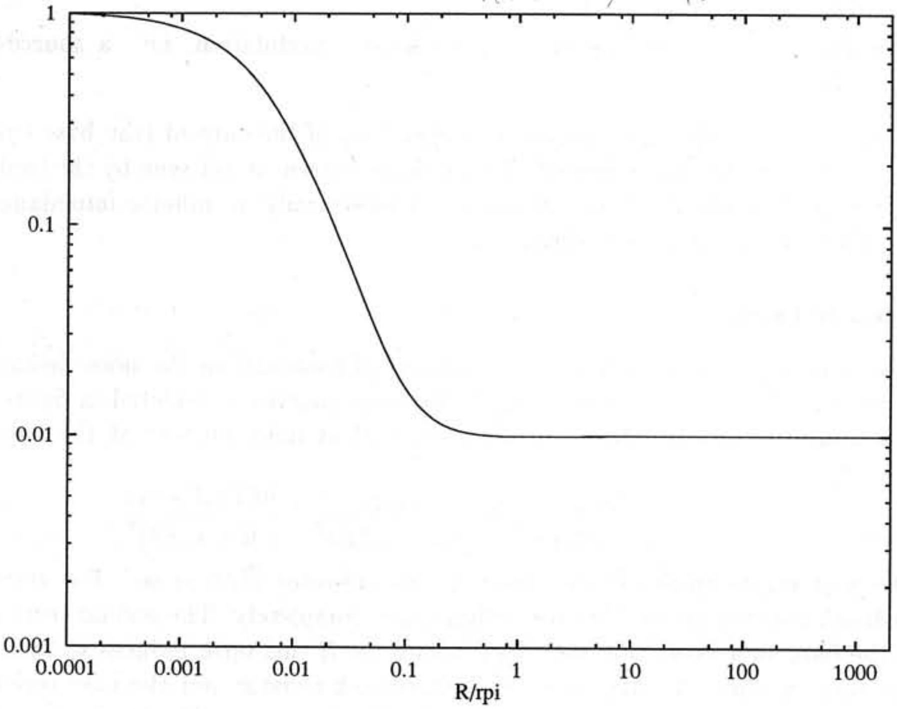


Figure 5.34: The noise of an active current source as function of the feedback resistor. The noise is normalized to $2qI_C$ and the feedback resistor to r_π and $\beta = 100$.

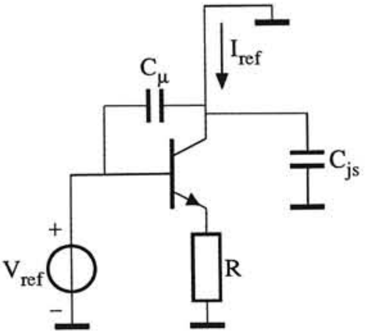


Figure 5.35: The parasitics of the active current source

impedance by negative feedback). A +3dB point, with respect to the minimal value, is obtained when the resistor is chosen such that the noise due to the resistor plus the noise due to collector current is equal to the base shot noise. In this case, the noise of the base resistance can, in general, be neglected. Then, R equals:

$$R = \frac{2\beta_f}{g_m}. \tag{5.86}$$

Mostly $\beta_f = 100$ and with $kT/q \approx 26\text{mV}$ the voltage across R equals:

$$V_R = I_C R = 2\beta \frac{kT}{q} \approx 5\text{V} \tag{5.87}$$

This is the same value as for the -3dB in the output impedance as a function of the feedback resistor. Apparently high-quality current sources need a supply voltage of minimally 5V.

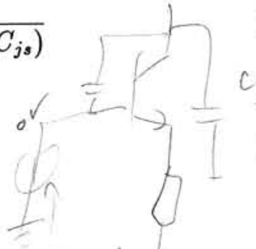
HF behavior

In the previous, sections only the dc behavior of the current source was discussed. However, that is only one part of the story. The current source has to behave well for high-frequency signals also, i.e. its output impedance has to remain relatively high. In figure 5.35, the source with its parasitic capacitances is depicted. The output impedance is given by:

$$r_{\text{out}}(j\omega) = \frac{r_{\text{out}}(0)}{1 + j\omega r_{\text{out}}(0)(C_\mu + C_{js})} \tag{5.88}$$

and has a pole at

$$p = \frac{-1}{r_{\text{out}}(0)(C_\mu + C_{js})}. \tag{5.89}$$



is a short - for high freq signal

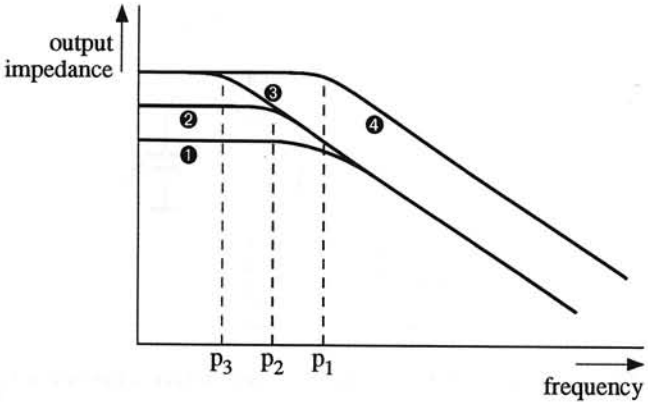


Figure 5.36: The effect of cascoding on the high-frequency output impedance: 1) without feedback and cascoding, 2) only feedback used, 3) feedback and cascoding is used, 4) as 3 but without substrate capacitors

For frequencies above $1/[r_{\text{out}}(0)(C_\mu + C_{js})]$ the output impedance is dominated by the parallel capacitance $(C_\mu + C_{js})$. The effectivity of cascoding, at relatively high frequencies, depends on the values of C_μ and C_{js}. Four situations are depicted in figure 5.36. Function 1 depicts the output impedance of a single CE stage, no feedback or cascoding is used. Function 2 represents the output impedance of the source when feedback is used. The low-frequency impedance is increased. The high-frequency impedance is not affected by the feedback. Both capacitors $(C_\mu + C_{js})$, are still in parallel with the high output impedance of the series stage (see figure 5.35). The current leaking away through the two capacitors is not "seen" by the feedback resistors and, consequently, the feedback loop cannot suppress these currents. Function 3 depicts the output impedance of the source when cascoding is also used. The low-frequency output impedance is increased further. The high-frequency output impedance is again not affected. A substrate capacitor is still in parallel with the output impedance of the source (It is assumed that C_{js} is much larger than C_μ). Function 4 depicts the output impedance of a source using feedback and cascoding when no substrate capacitance is present. The high-frequency impedance is drastically increased. All the parasitic capacitors are part of the feedback loop and thus their influence is suppressed.

It may be clear that negative feedback and cascoding does not decrease the *influence* of substrate capacitors. However, when the *capacitance* of the parasitics is decreased, some profit can be obtained. When a PNP transistor is used instead of the NPN, the sign of the current is changed but the output capacitance is decreased, because the substrate capacitor is connected to the base of the PNP and thus cascoding may help.

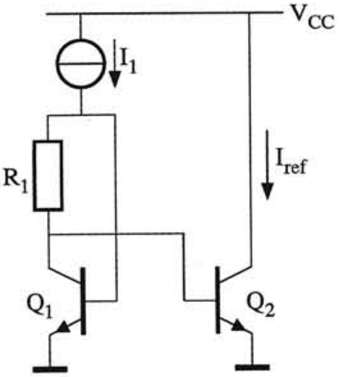


Figure 5.37: The peaking current source

Still, when all the measures are taken, the output impedance may be too low at (very) high frequencies. A very straightforward method used in HF design is putting a resistor in series with the current source. The output impedance is now, at high frequencies, dominated by this resistor instead of by the parasitic capacitance.

5.4.2 The peaking current source

The peaking current source is a special type of current source. The circuit is shown in figure 5.37. The relation between I_1 and I_{ref} is given by:

$$\ln\left(\frac{I_{\text{ref}}}{I_{S2}}\right) = \ln\left(\frac{I_1}{I_{S1}}\right) - \frac{I_1 R_1}{\frac{kT}{q}} \tag{5.90}$$

or

$$I_{\text{ref}} = I_1 \left(\frac{I_{S2}}{I_{S1}}\right) \exp\left(-\frac{I_1 R_1}{\frac{kT}{q}}\right) \tag{5.91}$$

with I_{S1} and I_{S2} the saturation currents of Q_1 and Q_2, respectively. The function is depicted in figure 5.38 with $\frac{kT}{q} = 26\text{mV}$, $I_{S1}/I_{S2} = 1$ and $R_1 = 10\text{k}\Omega$. This function exhibits a peak, which explains the name "peaking current source". At this extreme, a deviation in I_1 is not transferred to I_2. The source is biased at this extreme when:

$$RI_1 = \frac{kT}{q}. \tag{5.92}$$

When the voltage across R is equal to the thermal voltage, a change in the current I_1 is totally suppressed. No change in the output current is seen. At this extreme,

$R\ 500\mu = 25m$

$R = \frac{25m}{500\mu} = \frac{25,0}{500}\ 10^{-3} = \frac{1}{2}\ 10^{2} = 50\ \Omega$

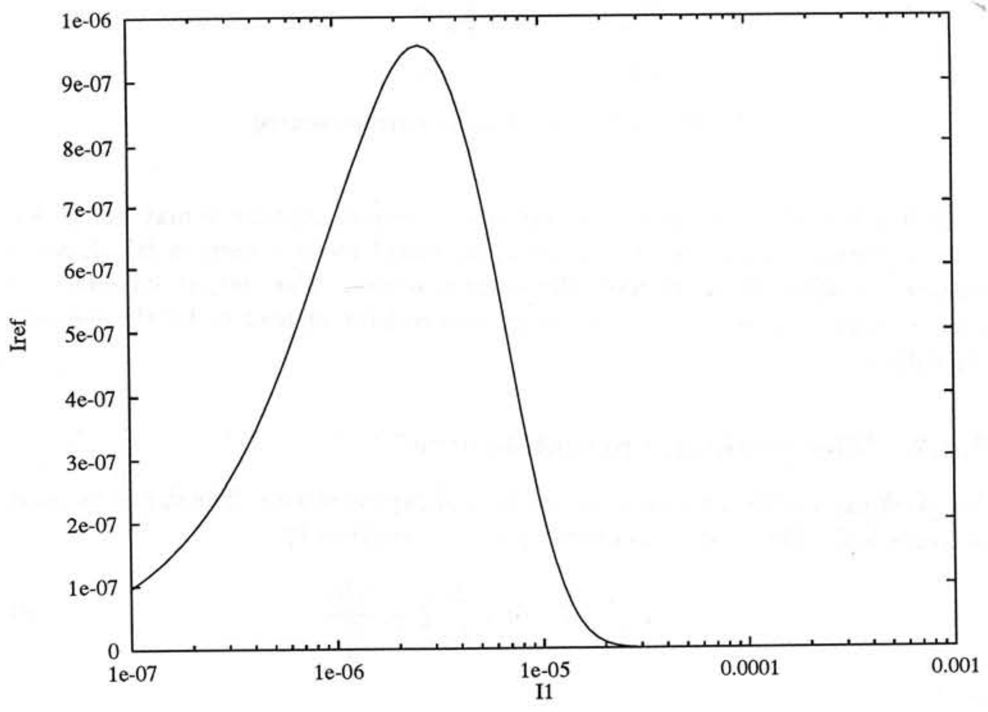


Figure 5.38: The relation of the input and output current (I_1 and I_{ref}) of the peaking current source

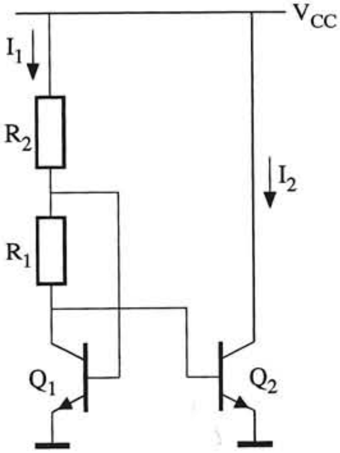


Figure 5.39: An example of a peaking current source

the ratio between the input and output current is given by:

$$\frac{I_2}{I_1} = \frac{1}{\mathrm{e}}\left(\frac{I_{S1}}{I_{S2}}\right). \tag{5.93}$$

An example of a peaking current source of 100μA is given in figure 5.39. The current in the left branch, I_1, must be 272μA. At 300K, R_1 must have a value of 95Ω (5.92). The current source on top of the left branch (figure 5.37) is implemented by resistor R_2 and must have a value of:

$$R_2 = \frac{V_{\mathrm{CC}} - V_{BE}}{272\mu\mathrm{A}} \approx 15.8\mathrm{k}\Omega \tag{5.94}$$

in the case of $V_{\mathrm{CC}} = 5\mathrm{V}$ and $V_{BE} = 0.7\mathrm{V}$.

Only resistor R_1 needs to be accurate. This poses no problem because it has a rather low value. On chip, a relatively wide resistor can be used. For resistor R_2, it is not necessary to be accurate because small changes in R_2 can be seen as small changes in I_1 and these are not transferred to I_2. For R_2, a rather thin resistor can be used.

As a consequence of the suppressing of small changes in I_1, this type of current source exhibits a very high PSRR. Even better results can be obtained by the enhanced peaking current source. This source is depicted in figure 5.40. The current mirror (to be discussed in section 5.5) at the top of the source forces the ratio of the two branch currents to be constant.

The non-ideality which has the most dominant influence on the performance of this source is the Early effect. When this effect is taken into account, the optimum

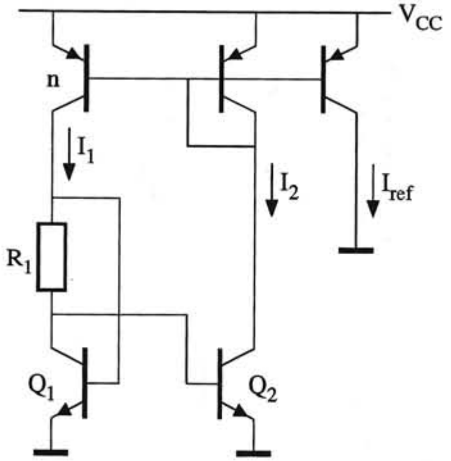


Figure 5.40: The enhanced peaking current source

bias condition is changed. The optimal ratio of the two branch currents equals in this case:

$$\frac{I_2}{I_1} = \exp\left(-\left[1+\frac{V_{AF-pnp}}{V_{AF-npn}}\right]\left[1-\frac{2V_{CE,\mathrm{pnp}}}{V_{AF,\mathrm{pnp}}}\right]\right) \tag{5.95}$$

with $V_{AF,\mathrm{npn}}$ and $V_{AF,\mathrm{pnp}}$ the forward Early voltages of the npn and pnp transistor, respectively. A current variation of only 0.05% over a supply voltage range of 10V can be obtained with this source.

5.5 The current mirror

The current mirror is a rather general-purpose circuit. The current mirror can be used as:

- current inverter/amplifier,
- bias source,
- translinear circuit.

The current mirror is usually used as a current source. The mirror when used as a translinear circuit is discussed in another chapter. The basic current mirror, based upon two bipolar transistors, is given in figure 5.41. The nullor forces the input current to flow completely through the collector terminal of Q_1 and the output transistor mirrors that current. When the output transistor is chosen to

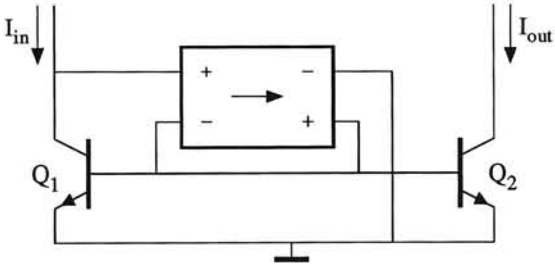


Figure 5.41: The basic current mirror

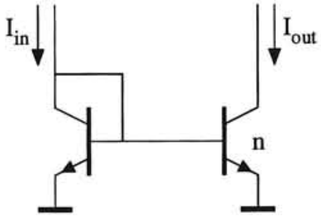


Figure 5.42: The simplest current mirror

be n times larger than the input transistor, and the Early effect of the transistor is negligible, the output current equals n times the input current:

$$I_{\text{out}} = nI_{\text{in}}. \tag{5.96}$$

By choosing the input transistor to be larger than the output transistor, n can be made smaller than one.

The simplest current mirror is depicted in figure 5.42. In this current mirror, the nullor is implemented by just a wire. Due to several non-idealities, the ratio of the input and output current is not exactly equal to n.

5.5.1 Errors in the mirror factor

There are three main sources of error in the current mirror shown in figure 5.42, namely:

- the base current (finite β),
- mismatch in the emitter areas,
- the Early effect.

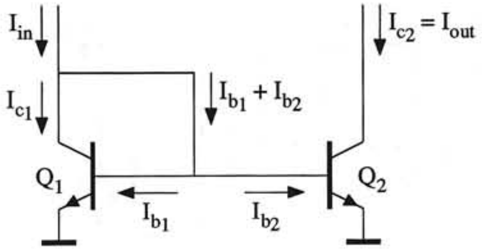


Figure 5.43: The simplest current mirror with all the currents flowing in the circuit depicted

Errors due to the base currents

In figure 5.43, the simplest current mirror is again depicted but now with all the currents flowing in the mirror and a scaling factor of 1. The output current is equal to the collector current of the input transistor and is equal to:

$$I_{\text{out}} = I_{C1} = I_{\text{in}} - \frac{I_{C1}}{\beta_1} - \frac{I_{C2}}{\beta_2}. \tag{5.97}$$

Consequently, the transfer equals:

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \frac{1}{1+2/\beta}. \tag{5.98}$$

The higher the current-gain factor is, the smaller the difference between the input and output current and the closer the transfer approaches 1. The relative error is:

$$\frac{I_{\text{in}} - I_{\text{out}}}{I_{\text{in}}} = \frac{1}{1+\frac{1}{2}\beta}. \tag{5.99}$$

It is assumed that the transistors have equal current-gain factors. The influence of the base currents can be minimized by using a better implementation for the nullor shown in figure 5.41. This is depicted in figure 5.44. The nullor is now implemented by a CE stage. For the input and output currents holds:

$$I_{\text{in}} = I_{C1} + \frac{I_{C1} + I_{C2}}{\beta^2} \tag{5.100}$$

and

$$I_{\text{out}} = I_{C1} \tag{5.101}$$

again with equal emitter areas. The relative error is given by:

$$\frac{I_{\text{in}} - I_{\text{out}}}{I_{\text{in}}} = \frac{1}{1+\frac{1}{2}\beta^2}. \tag{5.102}$$

The influence of the base current is decreased by a factor β. This equals the increase of loop gain in the mirror due to Q_3.

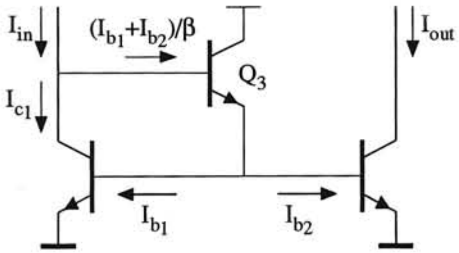


Figure 5.44: A current mirror with a reduced influence of the base currents

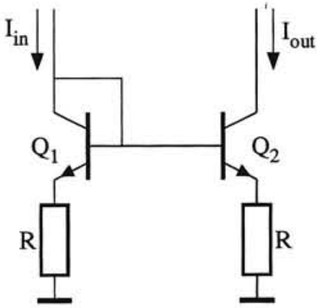


Figure 5.45: A current mirror with emitter resistors

Errors due to mismatch in emitter areas

In practical implementations, the two emitter areas are never exactly the same size. There is always a mismatch between them. Because of the linear dependence of the saturation current of the transistor on the emitter area, the two saturation currents have the same mismatch. Assume the two emitter areas, A_1 and A_2, relate to each other as:

$$A_1 = (1 + \Delta)A_2 \tag{5.103}$$

with Δ the relative matching error. The relative error of the input and output current is then given by:

$$\frac{I_{\text{in}} - I_{\text{out}}}{I_{\text{in}}} = \frac{A_1 - A_2}{A_1} = \frac{\Delta}{1 + \Delta} \approx \Delta. \tag{5.104}$$

It is assumed that the matching is not too bad, thus $\Delta \ll 1$. The relative error in the mirror factor is equal to the relative error in the emitter areas.

The error due to mismatch can be reduced by the use of emitter resistors. This is depicted in figure 5.45. Now the mirror factor is not only determined by the transistor areas but by the emitter resistors also. For the maze of this mirror holds:

$$I_{\text{in}}R + V_{BE1} = I_{\text{out}}R + V_{BE2}. \tag{5.105}$$

$I_{ab} = I_s e^{V_{be}/V_T}$

The relative error in the mirror factor, Δ', for the current mirror using emitter resistors is defined as:

$$\Delta' = \frac{I_{\text{in}} - I_{\text{out}}}{I_{\text{in}}}. \tag{5.106}$$

Combining equations (5.105) and (5.106) results in:

$$\begin{aligned}
\frac{(I_{\text{in}} - I_{\text{out}})R}{\frac{kT}{q}} &= \ln\left(\frac{I_{\text{out}}}{I_{\text{in}}} \cdot \frac{A_1}{A_2}\right) && (5.107)\\
&= \ln(1-\Delta') - \ln(1-\Delta) && (5.108)\\
&\approx \Delta - \Delta' && (5.109)\\
&. && (5.110)
\end{aligned}$$

$= V_{be_2} - V_{be_1}$

$= \ln \frac{I_{ou}}{I_s} \ln \ln \frac{I_{i1}}{I_s}$

Rewriting for Δ' yields:

$$\Delta' = \Delta\left(\frac{1}{1 + \frac{I_{\text{in}}R}{kT/q}}\right). \tag{5.111}$$

The mismatch error is approximately reduced by a factor equal to the ratio of the voltage across the emitter resistor and the thermal voltage. As was seen in the previous section, the emitter resistors also reduce the noise level and increase the output impedance of the mirror. A drawback is the required higher power-supply voltage. The use of emitter resistors improves the performance of the mirror considerably, but, for optimal performance, more than 5V supply voltage is required.

Of course, the matching of the resistors is important. For large voltages across the resistors, the transistor mismatch becomes negligible, only the mismatch of the resistors remains. However, usually, the matching of resistors is better than the matching of transistors because they are larger.

Errors due to the Early effect

The influence of the Early effect on the mirror factor is depicted in figure 5.46. The input current equals 1mA, and the output current varies between approximately 1 and 1.3mA. It is clear that major errors can occur due to the Early effect, i.e. the output impedance, of the transistor. There are two solutions for this problem. They can be used simultaneously:

- Use emitter resistors to increase the output impedance;
- Use a voltage follower which makes the two collector voltages equal.

The mirror with reduced Early effect is depicted in figure 5.47. Because of the emitter resistors the influence of the Early effect reduces, and the output

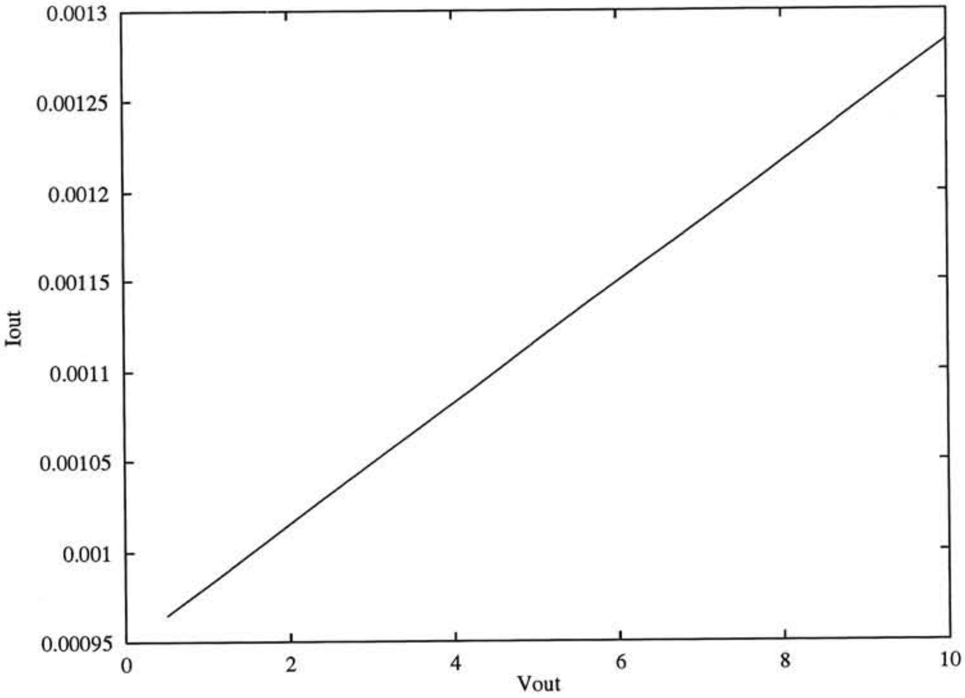


Figure 5.46: The influence of the output voltage on the output current

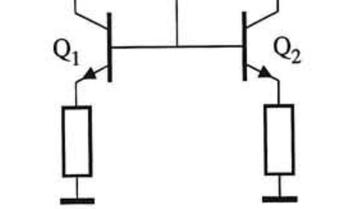


Figure 5.47: A current mirror with a reduced Early effect

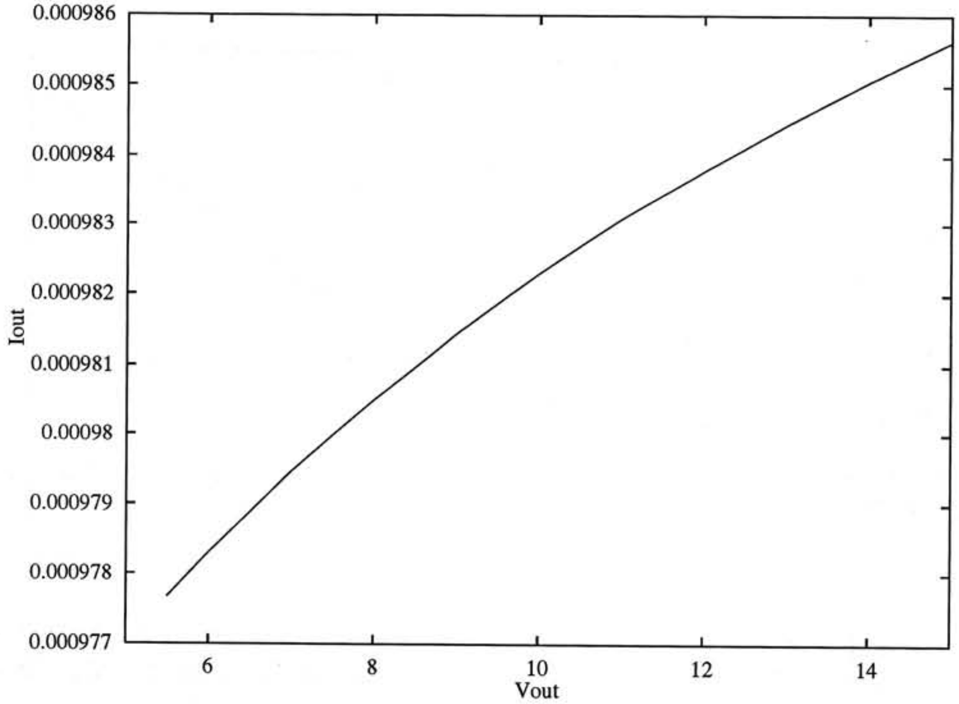


Figure 5.48: The output current as a function of the output voltage

impedance increases, as discussed in section 5.4.1. A further reduction of the error is attained by ensuring the equality of the two base-collector voltages. Q_3 is used as a voltage follower. The difference between the input and output voltage is one base-emitter voltage. The second pnp transistor, Q_4, is used as a level shift of one base-emitter voltage to compensate for this remaining difference. Now both base-collector voltages are equal and the error due to the Early effect is reduced even further.

The influence of the base currents is the same as it is in the simplest current mirror. Because PNP Q_3 functions as a current follower for the base currents of Q_1 and Q_2, the difference of the input current and the collector current I_{C1} is still the two base currents of Q_1 and Q_2. The output current of the mirror as a function of the output voltage is depicted in figure 5.48. Again an input current of 1mA was chosen. The influence of the output voltage on the mirror factor is decreased dramatically.

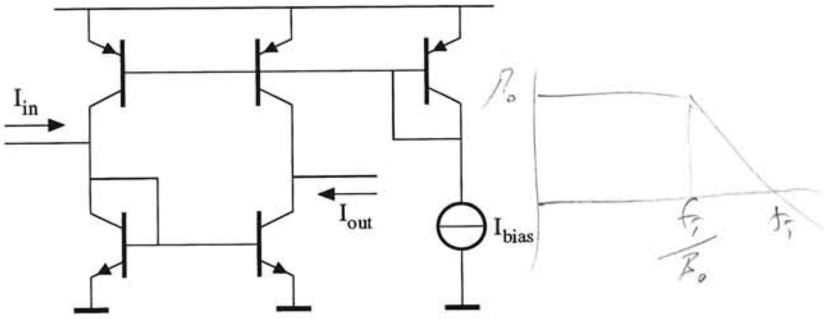


Figure 5.49: A PNP-like mirror with the frequency behavior of NPNs

High-frequency behavior

To examine the high-frequency behavior of the current mirror, the current-gain factor of the transistors, is described by:

$$\beta_f = \beta_0 \cdot \frac{1}{1 + j\omega\tau_f\beta_0}. \qquad (5.112)$$

The pole at $-1/\beta_0\tau_f$ represents the finite bandwidth of the transistors. Substitution of this expression for β in the expression for the transfer of the current mirror (5.98) results in:

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \frac{1}{1 + 2/\beta_0} \cdot \frac{1}{1 + 2j\omega\tau_f}. \qquad (5.113)$$

The high-frequency behavior is given by a pole at half the transit frequency of the transistor. Of course, the substrate capacitance and the collector bulk resistance have influence on the high-frequency behavior also.

In bipolar processes, the npn characteristics are optimized and the pnp transistor is mostly a lateral transistor with minor HF behavior with respect to the npn. For example in the DIMES-01 process:

pnp: $\tau_f = 8ns \rightarrow f_t = 20\text{MHz}$,

npn: $\tau_f = 30ps \rightarrow f_t = 5\text{GHz}$.

Sometimes in a design, a PNP mirror seems to be needed in the signal path. In such a case, the current mirror would introduce a pole at 10Mhz in the signal path. This can be a very inconvenient pole. However, it is possible to make a PNP-like mirror with the frequency behavior of an NPN mirror. The mirror is depicted in figure 5.49. The PNP mirror is taken away from the signal path and only used for biasing an NPN mirror. The NPN mirror performs the mirror function for the signal. The behavior of the PNP mirror is only important in respect to dc. Of course, the output impedance of the PNP mirror needs to be high in relation to the NPN mirror. The pole of this mirror is now at 2.5GHz.

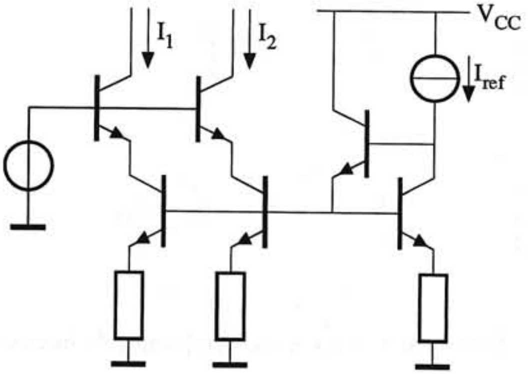


Figure 5.50: A current mirror extended to a current copier

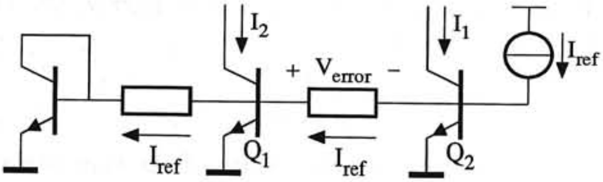


Figure 5.51: Errors due to incorrect wiring

5.5.2 Copying currents

When the current mirror is extended with additional output stages, a current copier is obtained. This copier is depicted in figure 5.50. The matching between I_1 and I_2 is equal to the matching of the emitter resistors. This can be very good. The matching between the input and output currents is as it is for the current mirror. Again the cascoding is needed for the reduction of the Early effect. When these current copiers are realized, care has to be taken with the connection of the base and emitter terminals. If there is a voltage drop across the wiring the effective base-emitter voltages can differ from each other, as depicted in figure 5.51. In the copier the reference current is connected far away from the input transistor. The reference current has to flow through the base terminal wiring. A voltage drop V_{error} of only 2mV between the base terminals of Q_1 and Q_2 already results in an error of 10%. The reference current should be supplied via a separate wire.

5.5.3 The MOS current mirror

In the previous sections, the current mirrors discussed were built with bipolar transistors, but they can be implemented with MOS transistors also. The MOS current mirror is depicted in figure 5.52. The transistors must always be in the

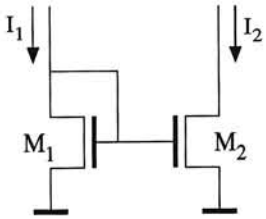


Figure 5.52: The MOS current mirror

pinch-off region. The gate-source voltage is higher than the threshold voltage (normally-off device) and the gate-drain voltage is zero.

The behavior of this current mirror is analogous to its bipolar equivalent. Errors due to the mismatch of transistor areas and parameters are mostly larger than for the bipolar equivalent due to the larger spread in parameters. The error due to the channel-length modulation ("Early effect") is generally larger too, because the output impedance of MOS transistors is lower, making cascoding favorable. In contrast with bipolar, the MOS mirror does not suffer from errors due to gate currents.

5.6 Self-biasing sources

Sources which need no additional sources for biasing purposes are called self-biasing sources. The value of the output signal and the biasing currents and voltages are referred to a internal voltage.

5.6.1 A self-biasing MOS current source

A very simple self-biasing source is depicted in figure 5.53. This source uses a

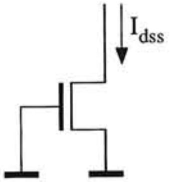


Figure 5.53: A very simple self-biasing current source

normally-on MOS transistor. The reference to which the output current is referred is an internal voltage. No additional reference is needed for the biasing. For a drain voltage higher than the threshold voltage, the MOS transistor is in the saturation region. In that region the current is given by:

$$I_d = \frac{\mu C_{\mathrm{ox}}}{2} \frac{W}{L} \left(V_{GS} - V_{TO}\right)^2 . \tag{5.114}$$

For the transistor shown in figure 5.53, the gate-source voltage is zero and consequently the output current is equal to the maximal saturation current:

$$I_{dss} = \frac{\mu C_{\mathrm{ox}}}{2} \frac{W}{L} \left(V_{TO}\right)^2 . \tag{5.115}$$

The MOS transistor suffers from a relatively low output impedance. To reduce the effect of channel-length modulation, a transistor with a relatively long channel needs to be used. To sink a large current the channel needs to be wide too, which results in a rather large MOS transistor.

5.6.2 A self-biasing PTAT current source

When a combination of a linear and a non-linear mirror is used, a self-biasing current source can be obtained also. This is depicted in figure 5.54. The linear mirror forces a linear relation between the two branch currents and the non-linear mirror a non-linear relation. The stable situation is the one for which both relation are fulfilled. These stable situations are given in figure 5.55 by the intersection points of the two transfer functions of the mirrors. There are two stable solutions: A trivial solution in which all the currents are zero, and the desired solution with currents unequal to zero. To force the circuit to the non-zero solution, additional measures have to be taken.

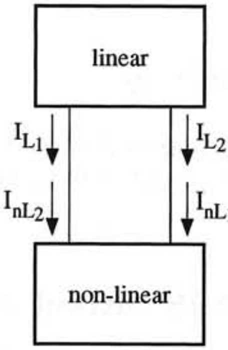


Figure 5.54: A self-biasing source using a linear and a non-linear mirror

An implementation of this type of source was given in the previous section, the enhanced peaking current source. Another widely used source is the PTAT current source. This source is given in figure 5.56. The source is just the PTAT

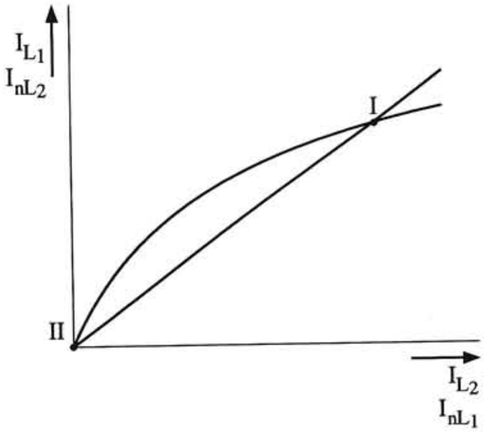


Figure 5.55: The solutions of the self-biasing current source using a linear and a non-linear current mirror

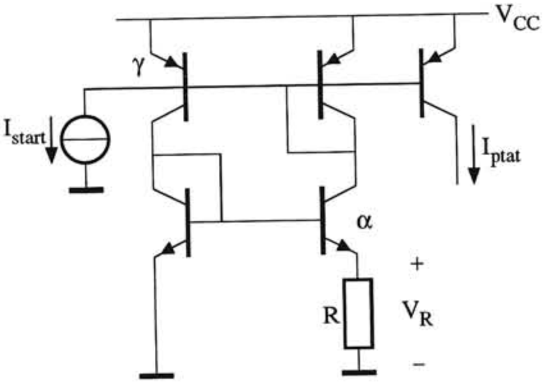


Figure 5.56: The PTAT current source

voltage source discussed in section 5.3.4, with the two current sources in figure 5.15 implemented by a current copier. Because the voltage across the resistor R is PTAT, the current through the resistor is PTAT also and this current is copied to the output. The output current is given by:

$$I_{\text{out}} = \frac{kT}{qR}\ln(\gamma\alpha) \tag{5.116}$$

with γ and α the scaling of the upper and lower mirror, respectively. To ensure the non-zero solution a start-up current is required. This current can be very low.

5.6.3 A self-biasing bandgap reference

A small extension to the source discussed in the previous section results in a self-biasing bandgap reference. The circuit is depicted in figure 5.57. Through the resistors R_1 and R_2 flows a PTAT current, I_{PTAT} and $(1+\gamma)I_{\text{PTAT}}$, respectively. The voltage across the two resistors is PTAT also and equals:

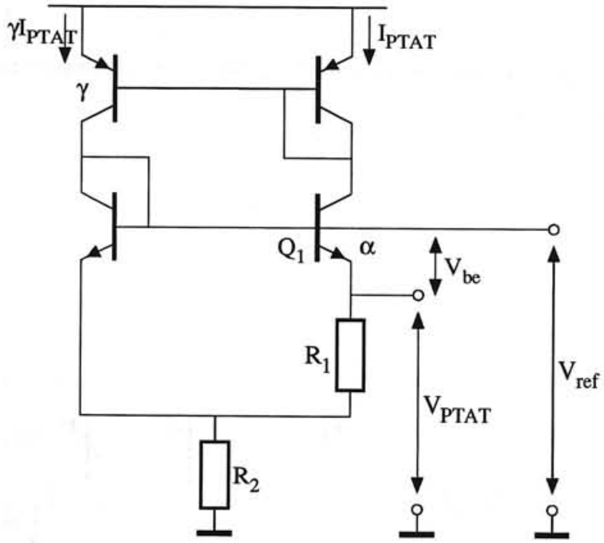


Figure 5.57: A self-biasing bandgap reference

$$V_{R_1+R_2} = V_{\text{PTAT}} = I_{\text{PTAT}}R_1 + (1+\gamma)I_{\text{PTAT}}R_2. \tag{5.117}$$

The base-emitter voltage of Q_1 is added to this PTAT voltage. By choosing the appropriate values for the resistors, the PTAT voltage compensates the first-order temperature behavior of the base-emitter voltage. This is just a bandgap reference of the type discussed in section 5.3.5. Although the quality is minor to the quality of the bandgap reference described in previous sections, this bandgap reference is a very simple one. Some errors in this bandgap are due to:

- base currents,
- errors in the mirror factor,
- the Early effect.

5.7 Reduction of the saturation voltage

The saturation of the transistor as discussed in section 5.3.4, is a desirable effect. For current sources, however, this is a highly disturbing effect. A single-transistor current source is depicted in figure 5.58. When the voltage V_{out} becomes too high the transistor goes into the saturation region. The base-collector junction becomes forward biased and the collector starts injecting charge carriers into the base region, resulting in a reduction of the output current.

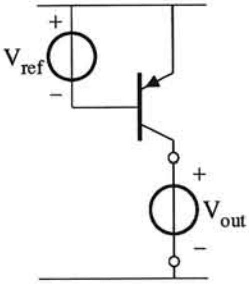


Figure 5.58: A single-transistor current source

Especially in low-voltage design, saturating current sources are a considerable problem. Because only a low supply voltage is available, i.e. 1V, the output voltage of a current source may not go beyond 0.8-0.9V. In such a case, the collector-emitter voltage remains higher than 0.1-0.2V and the source does not saturate. But the available supply voltage is reduced by 10-20%. To allow a larger output voltage swing, a reduction of the saturation voltage is favorable.

In section 5.3.4, an expression for the saturation voltage was given:

$$V_{CE} = \frac{kT}{q} \ln \left(\frac{1 + 1/\beta_r + \beta_{\text{sat}}/\beta_r}{1 - \beta_{\text{sat}}/\beta_f} \right). \tag{5.118}$$

The saturation voltage can be made small by a high reverse current-gain factor. For instance, the transistor from the example given in section 5.3.4 is changed to have a reverse beta of 6. The saturation voltage with $\beta_{\mathrm{sat}} = 20$ becomes:

$$V_{CE_{\mathrm{sat}}} = 44.7\mathrm{mV} \text{ @ } 300\mathrm{K}. \tag{5.119}$$

A reduction of 25%. The reverse current-gain factor is the key parameter in reducing the saturation voltage. The reverse current-gain factor is, besides by layout, determined by the doping levels. However, for a given process, the doping levels cannot be changed by the circuit designer.

The influence of the layout is depicted in figure 5.59 for a vertical transistor. In figure 5.59a, a tiny emitter is made in a large base and collector. When this transistor is biased in reverse mode the collector works as the reverse emitter. Many carriers injected from the 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 to the reverse collector than the carriers injected at point B. The chance that a carrier from point A reaches the reverse collector is lower, due to the higher chance of recombination, than the chance that a carrier from point B reaches the reverse collector. The transistor shown in figure 5.59b has a relatively large emitter. Now most carriers injected from the reverse emitter have to travel only a short distance, reducing the total recombination. As the base current is determined by the recombination in the base region, the latter transistor has a higher reverse current-gain factor than the former transistor. Therefore, to obtain a low saturation voltage, the collector and emitter areas have to overlap each other as much as possible.

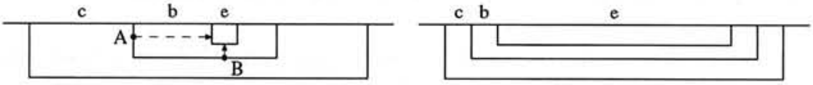


Figure 5.59: The influence of the layout on the saturation voltage. a) A transistor with a relatively tiny emitter and b) A transistor with a relatively large emitter

5.8 Conclusions on current sources

Several implementations of current constants have been treated above. The simplest current source uses a single resistor to derive the current from a voltage source. This source, however, requires a high supply voltage in the case of a high output impedance combined with a high output current. The active current source is able to realize a high output impedance and a high output current with a reasonable supply voltage. This source uses negative feedback. The feedback resistor determines the voltage-to-current conversion. When there is 5V across this resistor, noise performance and output impedance are close to the optimal values for this type of source.

The performance of the current mirror discussed thereafter, is shown to improve when emitter resistors are used: the output impedance increases, the mirror factor is closer to n, and the output noise level and the sensitivity for matching errors in the transistors reduce.

Finally, some self-biasing current sources were discussed. The advantage of this type of source is their simple structure. However, some of these sources may suffer from start-up problems.

Chapter 6

Analogue filters

Bert Monna

6.1 Introduction

The first filters ever made consisted of coils, capacitors and resistors. However, coils cannot be applied in integrated filters, thus the "conventional" design of continuous time filters does not fully cover the necessary design theory, especially not for active inductorless filters. Still, we give an historical introduction in order to highlight the development of 20th century filter design, because design theory developed for passive filters can be partially used for active filters. Alternatives implementations are presented for continuous time filters, and their pros and cons are discussed. An example of a passive filter, consisting of coils, capacitors and resistors is given. The example filter is transformed to a fully integrated continuous time equivalent filter. One of the most important design items in active integrated filters is the dynamic range, i.e. the noise level compared to the signal level present at the same moment. The computational labor required to determine the dynamic range is beyond the range of perspective, due to the elaborate matrix calculations, thus computers are required for calculations. The matrix computations are necessitated by the use of the state space description to describe the filter topology and transfer characteristic. The use of active components limits the dynamic range, and therefore the demands for optimization. The mathematical methods as well as the electronics are here presented.

Attention is paid to the implementation of integrators for continuous time filters in bipolar and in (bi-)CMOS technology. Finally, some practical problems in realizing continuous time filters are discussed.

6.2 History of filter design

Around 1890, several people were involved in improving the quality of transmission lines by adding coils. Only in 1899, did M.I. Pupin succeeded in improving the attenuation characteristic of telephone and telegraph wires by inserting coils. His success resulted in a world-wide use of "Pupin Lines". The behavior of these lines was more elaborately researched by G.A. Campbell, who, in 1903, published an article which described the frequency behavior of the lines. He invented the low-pass characteristic of the cable. He also realized the use of the cable as a band-pass filter, by replacing the coils by a combination of coils and capacitors. The problem of making filters with bulky cables led Campbell and K.W. Wagner to the simulation of the cable by a ladder construction of impedances. This was also indirectly suggested by Pupin, and the resulting filter was called the "electrical wave filter". The year 1915 may be considered the day of birth of the first electrical filter.

Design methods were invented by many people, amongst others by O.J. Zobel. The design method he developed was the beginning of the transmission line theory, that spoke in terms of characteristic impedance and wave propagation to describe the attenuation of the filter. He introduced a method to design filters with an infinite number of coils and capacitors. More filter theory was developed by S. Darlington and S. Butterworth. Butterworth made fourth-order filter sections, that were intercoupled by amplifiers (realized by tubes). Thus he was the first person to design active filters. Also from his hand are the well-known Maximally Flat Magnitude (MFM) attenuation characteristics (1930). Around the same time, W.R. Bennett solved the problem of realizing passive maximally flat transfer functions for filters of any order. W. Cauer also designed passive filters, but he used Tchebysheff approximations to describe the transfer function. Between 1930 and 1940, Cauer published several articles on the design of filters with some desired attenuation curve. In 1939, Darlington published an article in which he used Tchebysheff approximations to design transfer functions. The impact of Darlington's and Cauer's work was great, although the computing power in those days was too small to make full use of the theory.

Current monolithic technology does not allow the use of coils. The drawback of the generation of only poles on the real axis when making filters with only resistors and capacitors can be circumvented by using active components. Sallen and Key delivered a general design method to construct active R-C filters. It was based on cascading second-order stages. This method was not very popular in those days, because of the use of tubes. The emerging silicon technology, though, made it very attractive.

In 1977, the first switched-capacitor filter was applied. These filters still use a continuous signal amplitude, but process the signal at discrete time events. In

1979, Tan and Gray found solutions to tuning filters by placing automatic tuning circuits on chip. Tunability was realized by applying JFETs. Further important research was carried out by Moulding, Voorman, Tsividis, Nauta and Groenewold (tuning method, use of Gilbert Gain Cell for constructing integrators, MOSFET-C filters, high-frequency filters and dynamic range optimization, respectively).

6.3 Possible filter design

The drawback of making filters with coils and capacitors is the impossibility of integrating the coils. This resulted in the demand for fully integrated filters. Crystal and ceramic filters (both mechanical) are also commonly used filters. They usually have very high Q, do not need supply voltage, are cheap, but they cannot be integrated on chip, which is the major drawback. The previously mentioned Sallen and Key filters, which apply active components, can be used to design on-chip filters. They only use resistors, capacitors and transistors as active components.

Besides the digital filters, which apply mathematical operations on signals (no voltages or currents), the analogue active filters can be separated in two classes. The first class is that of the continuous time filters. Continuous time filters process the signal continuously in time, and use capacitors, resistors, coils and amplifiers to realize the filtering function on the basis of currents and voltages. The second class is the sampled-data class, consisting of the switched capacitor, switched current and switched voltage filters. The switched capacitor filters, for example, use switched capacitors to "simulate" resistors [1]. The capacitor is switched between the two connections, where the "resistor" should be. The charge transfer on the system clock signal, causes the capacitor to behave like a resistor. An integrator made in this way is depicted in fig.6.1.

The continuous time equivalent is depicted in fig.6.2. The equivalent resistor value yields:

$$R_s = \frac{1}{f_s C_r} \tag{6.1}$$

The resulting transfer function of the circuit yields:

$$\frac{1}{sR_s C_i} = \frac{f_s C_r}{sC_i} \tag{6.2}$$

The advantage of the last version is that no clock signal is necessary. Clock feed-through quite often causes the desired signal to deteriorate. What is more, the use of sampled data systems requires pre-filtering. The sampling causes higher

[1] This only describes half of what is possible. Using sampled data systems, it is also possible to make transfer functions that cannot be realized in continuous time systems, as for example FIR (Finite Impulse Response) filters.

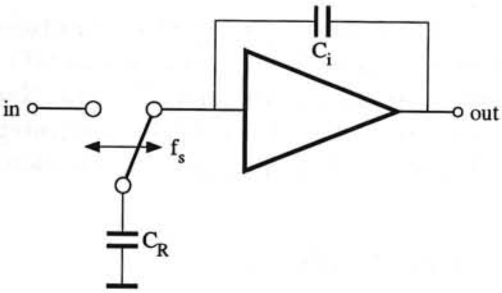


Figure 6.1: Example of switched capacitor (SC) integrator

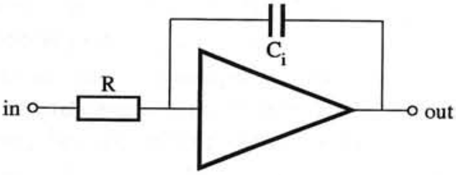


Figure 6.2: Equivalent continuous time (CT) integrator

frequency bands to be folded to the base band, i.e. aliasing. Thus, continuous time pre-filtering is required. The advantage of switched-capacitor filtering is its accuracy. The transfer function is fully determined by the matching of capacitors and an accurate clock signal. Usually, these requirements can be fulfilled. By changing the clock signal, the filter is tuned.

The dynamic range of continuous time and switched-capacitor filters appears to be the same for equivalent structures and high sampling rates. This can be explained because of the simulation of the resistor by a capacitor. On every clock pulse, some amount of charge is transferred to the following circuit part, depending on the value of the capacitor. Keeping in mind the aliasing problem, the capacitor can be viewed of as a resistor.

From this point, only CT filters are considered.

6.4 Overview of design trajectory

Filter design consists of several steps. The trajectory from specification to circuit is here discussed. Some of the steps will only be mentioned, because they are treated more thoroughly in following sections, others are merely mentioned to give a complete overview, but are actually beyond the scope of this book.

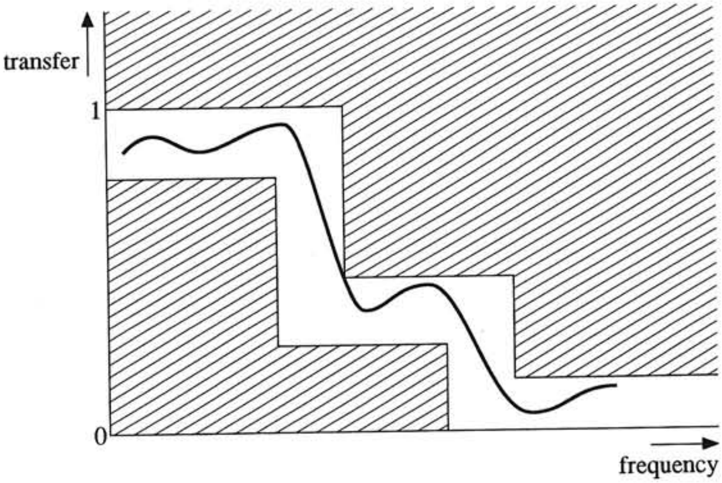


Figure 6.3: Example of attenuation curve

6.4.1 Specification of filter

The application in which the filter has to be used determines its specifications. The specification of the filter are several, for example: attenuation curve, dynamic range requirements, power consumption and supply voltage. The highest level specification must be handled first, which is the attenuation curve. In many applications, this attenuation behavior is given as depicted in fig.6.3. A filter curve must be designed such that the required attenuation curve is reached. It is usually attractive to use "standard" filter transfer functions. Well-known types are: Butterworth, Tchebysheff and elliptic or Cauer. The design of these filter types is usually based on a frequency normalized low-pass filter curve. Transformations are applied to the low-pass curve to turn it into, for example, a bandpass filter.

6.4.2 Frequency transformations

Almost any filter transfer can be derived from a low-pass equivalent filter. The transformation required is here given for the four most common filter structures. The transformations are based on the replacement of the original Laplace variable s by a new one. The normalized low-pass filter curve has its passband at lower frequencies and stop-band at higher frequencies, but the other filter types do not. The replacement of the original s by a "new" one, p, allows this p from 0 Hz to infinite frequencies to follow a contour along the original low-pass curve. For three often used filter types this is shown below.

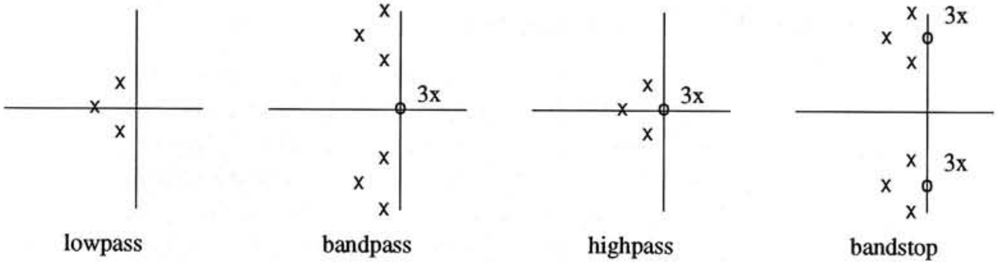


Figure 6.4: Pole zero diagrams of various filter types

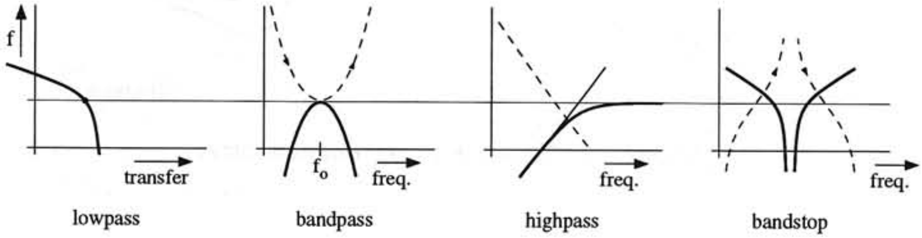


Figure 6.5: Visualization of frequency transforms

- low-pass to high-pass
  The transformation used is $s \leftarrow 1/p$. For low p (low frequencies), the s of the low-pass filter is large, which results in a large attenuation. For p large, the s is small, i.e. a passband behavior. These two regions result in a high-pass behavior.

- low-pass to bandpass
  The transformation used is $s \leftarrow p+p_0^2/p$. For small and for large p the equivalent s is large, which results in attenuation behavior. For p approximately around p_0, the equivalent s is in its passband. Therefore, this filter shows only passband behavior around p_0: a passband filter.

- low-pass to bandstop
  The transformation used is $s \leftarrow \frac{1}{p_0^2(p+1/p)}$. For low and high p, the equivalent s is low, thus resulting in passband behavior. For p around p_0, the corresponding s is high, therefore resulting in stop-band behavior.

The corresponding pole-zero pattern for a 3rd-order example system are shown in fig.6.4. The visualization of what actually happens when the low-pass equivalent s is replaced by the actual p variable is depicted in fig.6.5.

6.4.3 Mapping onto topologies

When the filter polynomial is known, or the pole and zero positions, which is actually the same, a topology should be found to map these information on. A topology is an idealization of a filter. It consists of ideal integrators (branches valued $1/s$) and an interconnection circuit. There are many ways to realize a filter function. Some possible topologies are shown in one of the following sections The choice of a topology is mostly based on experience of "good behavior". There are some topologies that are known for their low sensitivity and good dynamic range behavior.

6.4.4 Implementation

The steps discussed so far only resulted in abstract filter structures. The structures consist of ideal integrators and interconnection circuitry. They are still implementation independent. The last step is to find an implementation of the integrators, in bipolar, MOS or BiMOS technology. This is also discussed in the coming chapters.

6.5 Construction of continuous time filters

The most convenient way of designing continuous time filters is to use the known filter theory of passive filters. The design path followed uses standard filter tables to determine component values in order to attain the desired transfer function. The dynamic range -defined as the maximal signal with respect to the noise level that the filter is able to handle at the same time– is in the case of passive filters not limited by noise, because reactive components ideally do not introduce noise. The maximally possible signal levels at the capacitors and inductors are (almost) infinite; they are limited by the dissipation in the resistors and the "breakdown" in the capacitors and saturation of the inductors. Noise of the (also parasitic) resistors puts a lower limit on the smallest signals to be handled.

In active integrated filters, the poles and zeros are determined by resistors and capacitors, as well as by the active components. The use of resistors introduces noise, as does the use of active components. The active components are also assumed to operate within a certain supply voltage, thus limiting the maximal output signal. Hence, a translation is necessary to use the conventional filter theory for constructing active continuous time filters.

The above-mentioned problems necessitate the use of some other description, instead of only the transfer function. The state space description appears to be suitable. It not only establishes the desired transfer function, but also the topology of the filter, which appears to have a dominant influence on the dynamic range.

6.5.1 The state space description

The state space description is used to combine the transfer function and the topology of a filter in one description. As the filter can be viewed as a linear differential equation, the state space must be able to represent this. The variable s is the Laplace variable, which determines the poles and zeros of the filter. Suppose a transfer function of two polynomials (the order of the denominator is n, which is higher or equal to the order of the numerator). This transfer function can be represented in a signal flow graph as a connection of n integrators or differentiators. From this point, it is assumed that only integrators are used, as differentiators appear to be difficult to implement because of the critical stability considerations. This does not imply restrictions on the quality aspects of the filter or design freedom. The connection of the integrators can be described in the following equations:

$$sX = AX + BE_i \tag{6.3}$$

$$E_o = CX + DE_i \tag{6.4}$$

Thus, the new input of the integrators (sX)is a function of the old output signal of the integrators (X) and of the input signal (E_i). The output of the filter (E_o) is a combination of the output of the integrators and some fraction D of the input signal. The factor D can always be neglected in DR calculations, as this signal through this branch does not interfere with the internal structure of the filter.

A,B,C and D are matrices, which look like:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{6.5}$$

$$B = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \tag{6.6}$$

$$C = \begin{pmatrix} c_1 & \cdots & c_n \end{pmatrix} \tag{6.7}$$

$$D = (d) \tag{6.8}$$

The resulting transfer function is:

$$H(s) = \frac{E_o}{E_i} = C(sI - A)^{-1}B + D \tag{6.9}$$

This is rendered schematically in fig.6.6.

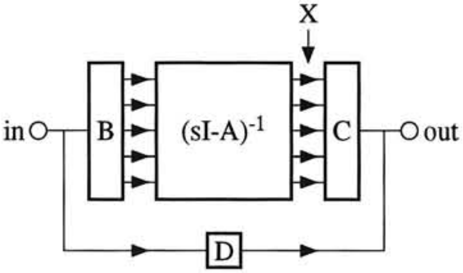


Figure 6.6: Scheme of state space description

6.5.2 Ladder realization from passive filter

There are some known methods to determine a topology and to realize a transfer function. One method is to derive the state space description from a passive filter, after which an active implementations is constructed.

There are two means of realizing active integrators. First the use of active integrators, i.e. an opamp with a capacitor in feedback, or second an active g_m, followed by an integrating capacitor. The design trajectories are shown in fig.6.7.

6.5.3 Cascade Realization

The cascade realization can be directly found by splitting the denominator into first- and second-order transfer functions. Applying this method, a third-order Butterworth filter (normalized) is written as:

$$H(s) = \frac{\frac{1}{2}}{(s+1)(s^2+s+1)} \tag{6.10}$$

This results in the signal flow graph shown in fig.6.8. By implementing the $1/s$ branches by active integrators, an active filter results.

6.5.4 Direct realization

The direct realization extracts the signal flow graph directly from the transfer function. Taking the Butterworth characteristic:

$$H(s) = \frac{\frac{\omega_c^3}{2}}{s^3 + 2\omega_c s^2 + 2\omega_c^2 s + \omega_c^3}. \tag{6.11}$$

This can also be written as:

$$H(s) = \frac{b_1}{s^3 - a_{33}s^2 - a_{23}s - a_{13}} \tag{6.12}$$

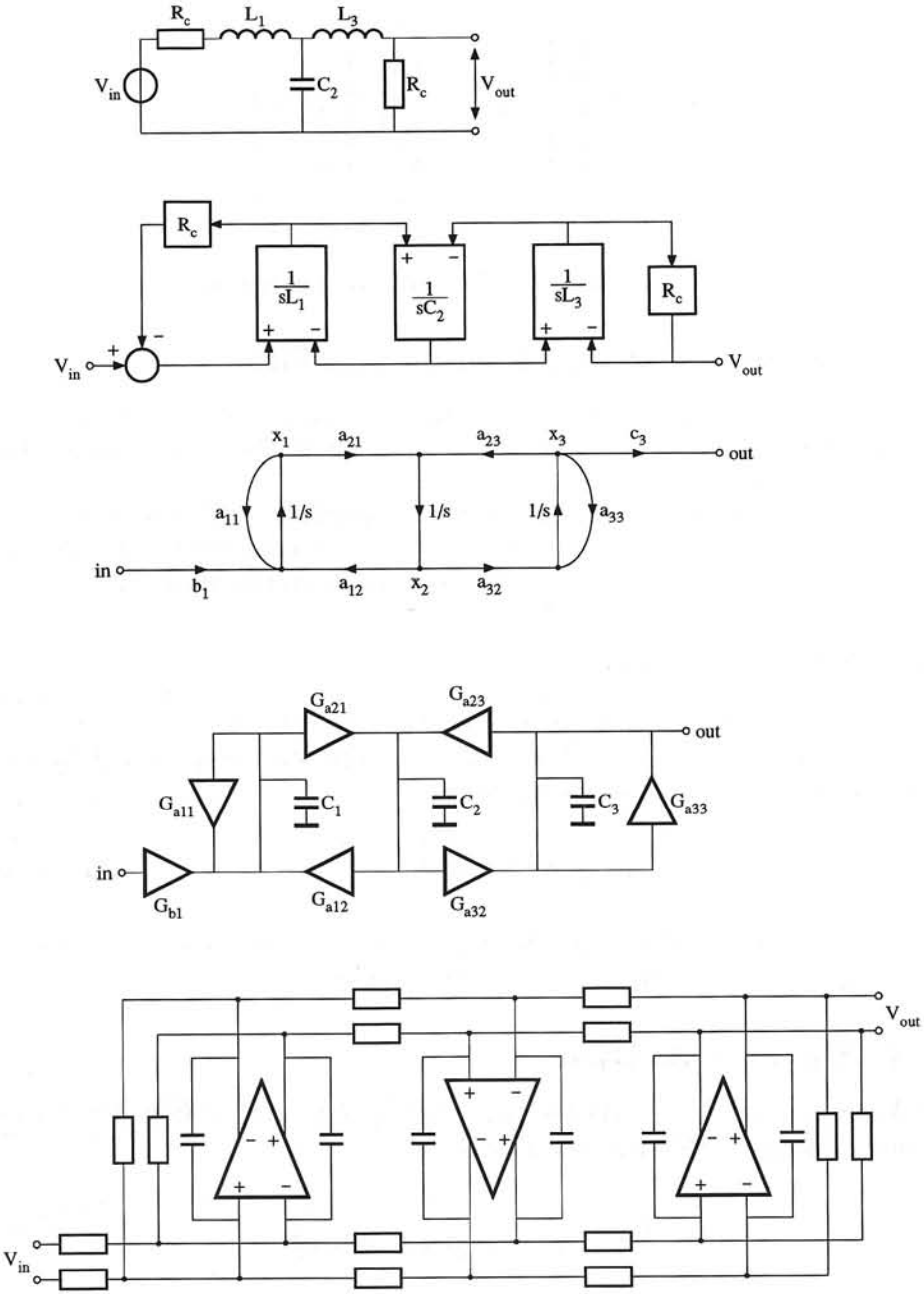


Figure 6.7: Transformation from passive filter to active filter

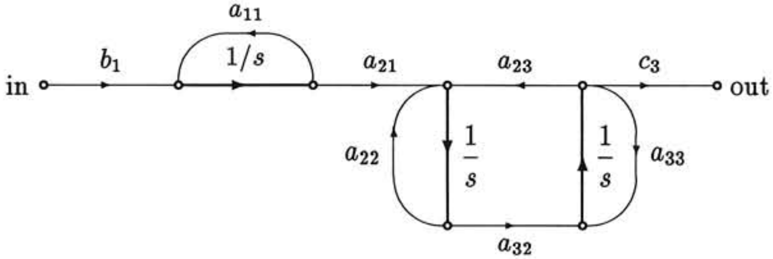


Figure 6.8: Signal flow graph of cascade realization

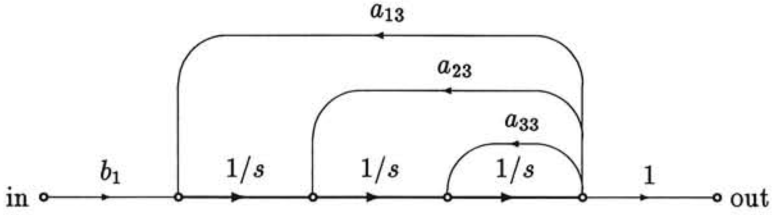


Figure 6.9: Signal flow graph of direct realization

This results in the signal flow graph shown in fig.6.9, from which an active filter can be derived.

The former examples of the implementation of the transfer functions show that it is possible to use several topologies to realize the same transfer function. This, however, does not imply that the various topologies behave the same. Various topologies, for example, appear to show different dynamic range behavior, but also different sensitivity behavior.

6.6 Integrators

Filters can be considered to be composed of integrators or differentiators. Integrators are used because differentiators are inherently difficult to implement, due to stability problems. In continuous time filters, using the Laplace domain description, the integrating function can be defined as the transfer a_0/s. There are two integrating elements in electronics, i.e. coils and capacitors. As high-quality coils cannot yet be integrated on a chip, capacitors almost always are used to implement the integrator function. The current that flows through a capacitor results in an integrated voltage across the capacitor terminals. If the coil were taken as an integrating element, the voltage across the coil would result in an integrated current through the coil.

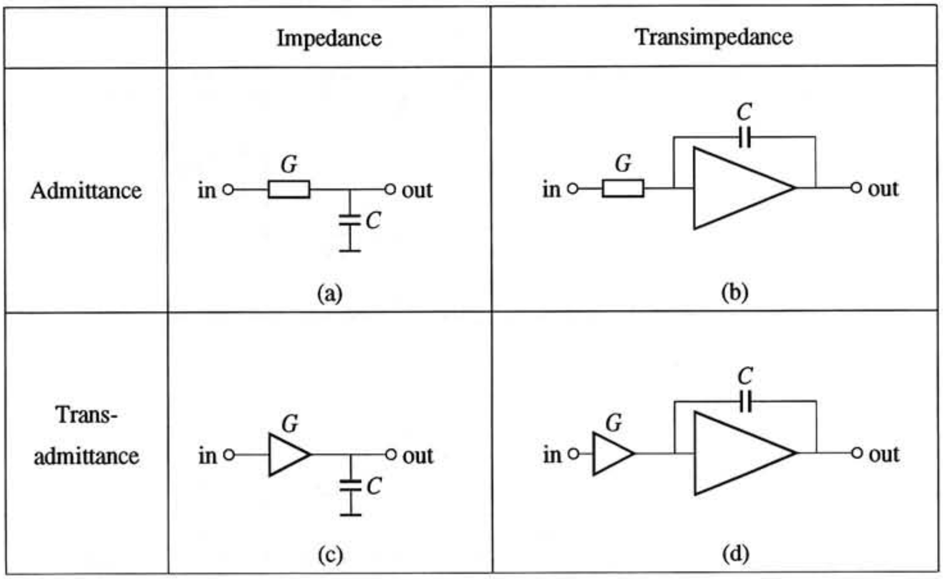


Figure 6.10: The four classes of integrators

Choosing the capacitor as the integrating element implies using a current as input quantity, and getting a voltage as output quantity. This necessitates using voltage-to-current conveyors, or (trans-)conductances in order be able to connect the integrators. The simplest solution is to make use of resistors.

Actually, four types of voltage-voltage integrators are possible. These are depicted in fig.6.10.

The four types differ in the way they implement the ideal integrator function, by making use of active components. The admittance-impedance integrator does not use active components. However, with this type of integrator, it is not possible to make filters with complex poles. Therefore, this type of integrator is not used.

The second type of integrator is the admittance-transimpedance integrator. In this type of integrator, the realization of the actual integration function is active. The advantage is that the "opamp" used is a well-known electronic function, that can be easily integrated. The opamp can be designed to operate rail-to-rail at the output terminals, so full advantage is taken of the supply voltage. This allows for optimal dynamic range performance. The resistor used can be integrated as a diffused resistor, but it could also be implemented as an MOS transistor in the triode region thus yielding the MOSFET-C filters.

The third type, the transadmittance-impedance integrator, makes use of active "resistors", or transconductances. The advantage of transconductors is that they are able to operate at high frequencies, because in these integrators the parasitic capacitors of the transconductor are in parallel with the time constant determining capacitors. In this case, they can be accounted for in the dimensioning of the required capacitance. A major drawback, however, is that it seems impossible to implement transconductors with rail-to-rail input capability.

The fourth type of integrator is the transadmittance-transimpedance integrator. This integrator has no advantages over the second and third integrators mentioned. The disadvantage is the use of two active parts. Both parts add distortion, as distortion is chiefly formed by active components and, moreover, the power consumption and the noise production increase.

Because of the above considerations, the second and third type of integrators are preferred.

6.6.1 Demands on integrators

As filters are composed of integrators, they rely on the properties of the integrators. It is not only the demands on dynamic range that are important, but also the demands on accuracy, tuning, etc. For all integrators it applies that process tolerances have to be dealt with, and in some applications filters have to be tuned. Here are two options: varying the capacitance, or varying the resistor. In order to design high-performance filters, the integrators have to have sound properties with respect to filter specifications.

One of the most important issues appears to be the dynamic range of the integrators. The dynamic range is defined as the ratio between the concurrent maximally possible signal level and the lowest possible signal level (equal to the noise level).

$$DR = \frac{V_{max}^2}{V_{noise}^2} \tag{6.13}$$

It can be clearly seen that the DR is not only dependent on the noise level, but also on the maximally possible signal level. Generally, it is easier to extend the output capability (V_{max}^2) of the integrator than the noise performance.

The implementing of integrators requires special attention be paid to the dynamic range. Implementations in various technologies are discussed in the following subsections.

6.6.2 MOS Integrators

MOS transistors in strong inversion are known to operate in two regions, i.e. the saturation and the triode regions. In addition, MOS transistors can be operated actively as well as passively. The difference is that actively operated MOS transistors have the input signal at the gate terminal, and passive transistors have the input signal at the source terminal. By making this division, four different types of

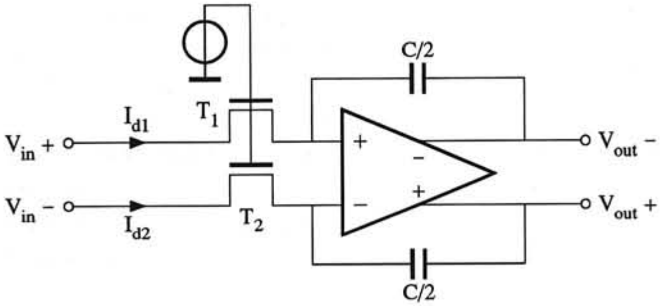


Figure 6.11: Passive triode integrator

MOS integrators are possible. In low-voltage applications, problems occur due to the threshold voltage of the MOS transistor. In some configurations, it is possible to use a charge pump to supply non-current consuming gate bias voltages.

Passive triode integrator

In the triode region, the MOS transistor behaves according to the following equation:

$$I_d = \frac{W}{L}\mu C'_{ox}\left[(V_{gs} - V_t)V_{ds} - \frac{1}{2}(1+\delta)V_{ds}^2\right] \tag{6.14}$$

This can be rewritten as:

$$\frac{I_d}{V_{ds}} = \frac{W}{L}\mu C'_o x\left[(V_{gs} - V_t) - \frac{1}{2}(1+\delta)V_{ds}\right] = f(V_{ds}) \tag{6.15}$$

δ is a bias-dependent parameter with a value of about 0.12. In the triode region, the $V_g > V_d + V_t$ and $V_g > V_s + V_t$ for NMOS transistors. From this equation, it becomes clear that it is possible to use the MOS transistor in the triode region for the voltage-to-current conversion, after which the current can be integrated into a voltage by a capacitor. The transconductance, however, is a function of the drain-source voltage. This implies non-linearity. Most of the even-order non-linearities can be eliminated by using balanced structures. An example can be seen in fig.6.11.

Active triode integrator

The active triode integrator is described by the same equations as the passive integrator. The only difference is the coupling of the signals to the transistor. The g_m of the MOSFET is used to couple the integrators. The distortion can also be reduced by applying balanced structures. An active triode integrator can be seen in fig.6.12.

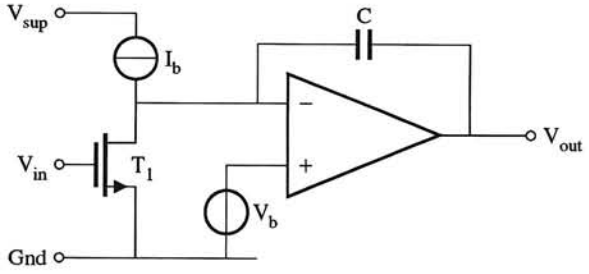


Figure 6.12: Active triode integrator

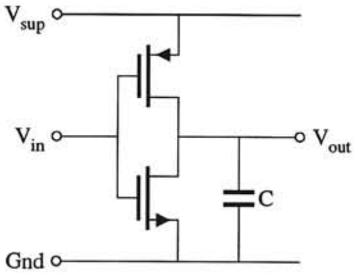


Figure 6.13: Active saturation integrator

Active saturation integrator

For MOS transistors in saturation, the following equation describes the relation between the drain current and the terminal voltages:

$$I_d = \frac{W}{L}\mu C'_{ox}\frac{(V_{gs} - V_t)^2}{2(1+\delta)} \tag{6.16}$$

In the saturation region $V_g < V_d$. Also in this integrator, even-order non-linearities can be canceled by using balanced structures. An example is given in fig.6.13.

This integrator has good high-frequency characteristics, because parasitic capacitances are in parallel with the desired capacitances, such that no parasitic poles occur. Because NMOS and PMOS transistor non-linearities partly cancel out, the distortion reduces but further measures are necessary to decrease the resulting distortion.

Passive saturation integrator

A passive saturation integrator uses the source terminal as the input and is biased in the saturation region. An example is given in fig.6.14.

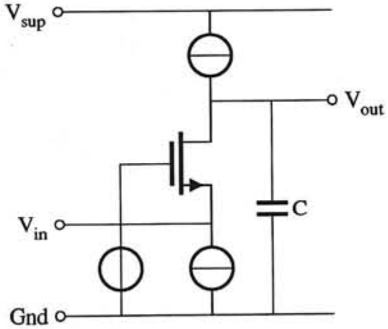


Figure 6.14: Passive saturation integrator

Reviewing MOS integrators

There are four types of MOS integrators. One of the transconductors has to be chosen according the application. Generally, the distortion can be diminished by using balancing. The triode integrators are followed by a virtually grounded integrator, because their transconductances depend on the V_{ds}, and thus is not allowed to vary. This is not the case with saturation integrators, so no use has to be made of opamp integrators.

Bipolar integrators

By using bipolar transistors, the g_m of the transistor can be employed to realize the desired voltage-to-current transfer. The other option is the use of the bipolar transistor as an amplifying element in feedback configuration. In this case, the actual voltage-to-current transfer is realized by another element. It is impossible to make use of the bipolar transistor as a resistor to the same extent as is possible with MOS transistors in the triode region. Tuning of such a filter is possible by varying the capacitance, as depicted in fig.6.15.

An example in which the g_m of a transistor is used to tune the filter is shown in fig.6.16. The resistor R allows for the voltage-to-current conversion. The current mirrors with input transistors Q_1 and Q_2 give a scaled transfer to the integrating capacitor, because of the different bias currents of the input and output transistors of current mirror transistors. This changes the differential resistance of Q_1 and Q_2, by which the coupling of the transistors changes. In both examples, some form of common-mode feedback is prerequisite to ensure a stable and desired bias voltage on the capacitor terminals.

A low-voltage integrator is depicted in fig.6.17. This integrator operates at 1-V and is of the "opamp RC" type. Bipolar transistors are used for high-gain balanced amplification. Diffused resistors and pinch resistors ensure the voltage-to-current

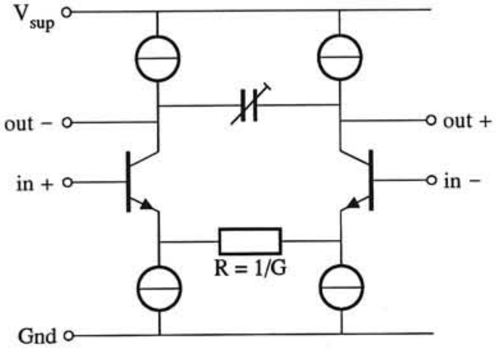


Figure 6.15: Bipolar integrator

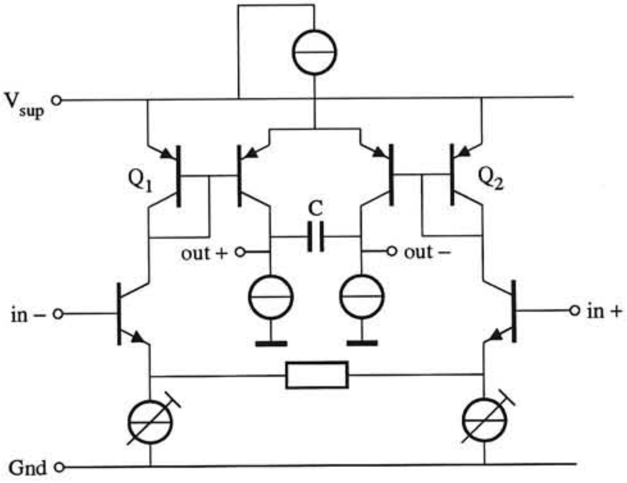


Figure 6.16: Bipolar integrator

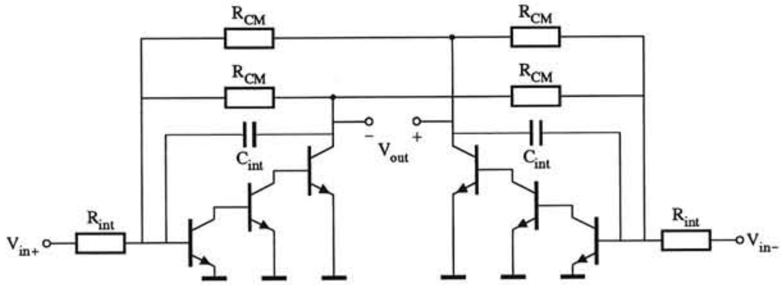


Figure 6.17: Bipolar integrator for 1-V applications

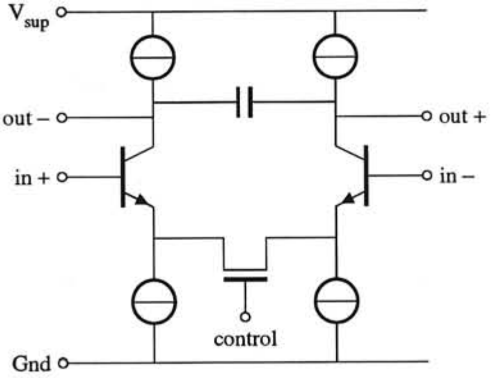


Figure 6.18: BiCMOS integrator (I)

transfer.

BiCMOS integrators

When BiCMOS is permitted, the advantages of both MOS and bipolar technologies can be used. Two examples are given in figs.6.18 and 6.19. The first one is the same as the bipolar integrator shown in fig.6.15. The tuning of the filter has been taken care of by implementing the resistor by means of an MOS transistor in the triode region. The other one is an MOS integrator of the active triode type. The integrator is tuned by the bipolar device. The bipolar transistor behaves like a voltage source, to ensure the drain voltage on the "kernel device", i.e. the device that actually fulfills the voltage-to-current transfer. By adapting this voltage, the integrator transfer changes.

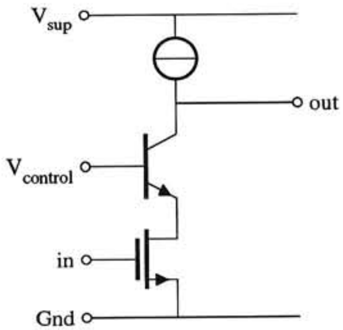


Figure 6.19: BiCMOS integrator (II)

6.7 Optimization of dynamic range

The use of coils and large capacitances has the disadvantage of not being integratable. The advantage of using coils and capacitors for realizing filters is that coils and capacitors have considerable signal-handling capabilities. The dynamic range of passive filters can in theory be infinite, because only coils and capacitors determine the location of the poles and zeros. The terminating resistors at the input and the output of the filter introduce some thermal noise, but this is very little.

The dynamic range of active filters is not infinite. Limitations arise at the two sides of the dynamic range. The output capability is limited by the supply voltage. No signals appear that go beyond the supply voltage. The noise level is introduced by the use of resistors to determine the filter transfer. The amount of noise becomes worse when active devices are used; this gives rise to a noise factor.

Although not evident, the dynamic range appears to be dependent on the topology of the filter. It can be proven that to every filter transfer there is some corresponding maximal dynamic range together with a given supply voltage, total capacitance and noise factor of the active devices. Only for one topology can it be proven that this maximum dynamic range can be reached. No attention is paid here to the exact calculations, we address only some clarification of the principles.

6.7.1 Matrix transforms

The state space description is suitable to describe the filter topology and transfer function. It is used to calculate new or adapted topologies in order to optimize the dynamic range.

There are two methods of optimization. The first one is called scaling. Scaling does not change the topology of the filter. Simple "start" topologies are locally optimized. This method does not usually reach the maximally possible dynamic range. Full optimization is, however, able to reach this limit. In this case the topology is completely changed. All possible connections between integrators may be used to obtain the maximal dynamic range.

The change of topology, without changing the transfer function, can be described by the following transformation with the transformation matrix T:

$$A' = T^{-1}AT \tag{6.17}$$

$$B' = T^{-1}B \tag{6.18}$$

$$C' = CT \tag{6.19}$$

$$D' = D \tag{6.20}$$

By which the states of the filters are transformed according to:

$$X' = T^{-1}X \tag{6.21}$$

By these transforms, the transfer function of the filter remains the same:

$$H'(s) = H(s) \tag{6.22}$$

Thus the transfer function remains the same, as the topology is changed, in order to attain the maximal dynamic range.

6.7.2 Dynamic range of integrators

The dynamic range of integrators consisting of a resistor and a capacitor (active or passive) is partly determined by the maximal signal levels the integrator is able to handle. Because integrators are coupled in a filter, all integrators must have approximately equal signal-handling capabilities. When the maximal signal amplitude is V_{max}, the maximal signal level is by definition $V_{max}^2/2$.

The noise of the integrator can be modeled as a noise voltage source at the input, with a double-sided spectrum of:

$$S_{ni}(\omega) = \frac{1}{2\pi}\frac{2kT\xi}{|G|} \tag{6.23}$$

in which ξ is the noise factor of the integrator with a minimal value of 1. This noise factor is the surplus noise of the active devices. The mean squared noise voltage can be determined by integrating over the noise bandwidth B, which can be chosen equal to the unity gain frequency of the integrator ($|G|/C$):

$$\overline{V_{ni}^2} = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_{ni}d\omega = \frac{1}{2\pi}\int_{-G/C}^{G/C} S_{ni}d\omega = \frac{2kT\xi}{\pi C} \tag{6.24}$$

From this it can be concluded that the noise level is dominantly determined by the capacitance. The dynamic range of the integrator becomes:

$$DR = \frac{\pi V_{max}^2 C}{4kT\xi} \tag{6.25}$$

A first-order filter (integrator) like the one shown in fig.6.2 has a noise voltage source which equals:

$$S_G(\omega) = \frac{1}{2\pi}\frac{2kT}{G} \tag{6.26}$$

The output noise spectrum becomes:

$$S_o(\omega) = \frac{1}{2\pi}\frac{2kT/G}{1+(\omega C/G)^2} \tag{6.27}$$

Integrating over the total frequency band, the total output voltage noise $\overline{V_{no}^2}$ appears to be kT/C. The dynamic range becomes, by a maximal signal level of $\frac{1}{2}V_{max}^2$:

$$DR = \frac{V_{max}^2 C}{2kT} \tag{6.28}$$

The difference between this formula for the DR and the former is the difference constituted by the definition of bandwidth. In the last situation the bandwidth is known, because the filter network is known. If the integrator is not placed within the context of the filter network, the noise bandwidth is unknown, so an assumption has to be made. This assumption does not have any influence on the optimization of the dynamic range, only on the absolute value of the dynamic range that results.

Optimization of the dynamic range of filters

Two items are important when optimizing filters: the maximal signal capability and the noise. The problem can be viewed as that of looking through a window. This is depicted in fig.6.20 . All windows have a view which extends from left to right; it can be large or small. The total range is determined by the highest noise level of all integrators, and the lowest output level of all integrators. Although all integrators can have a large dynamic range, not all integrators make optimal use of it. Better performance results when all the integrators are scaled to each other. The original output levels of the integrators of a third-order filter are depicted in fig.6.21. It can be seen that the first integrator has to handle larger signals than the second and the third integrator. By scaling the top levels of the integrator's output levels, the configuration shown in fig.6.22 results. This kind of optimization is especially suited for sinusoidal input signals.

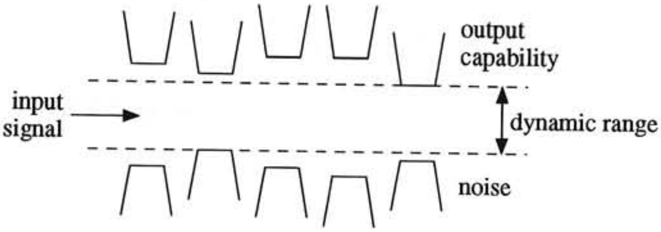


Figure 6.20: Window of several integrators

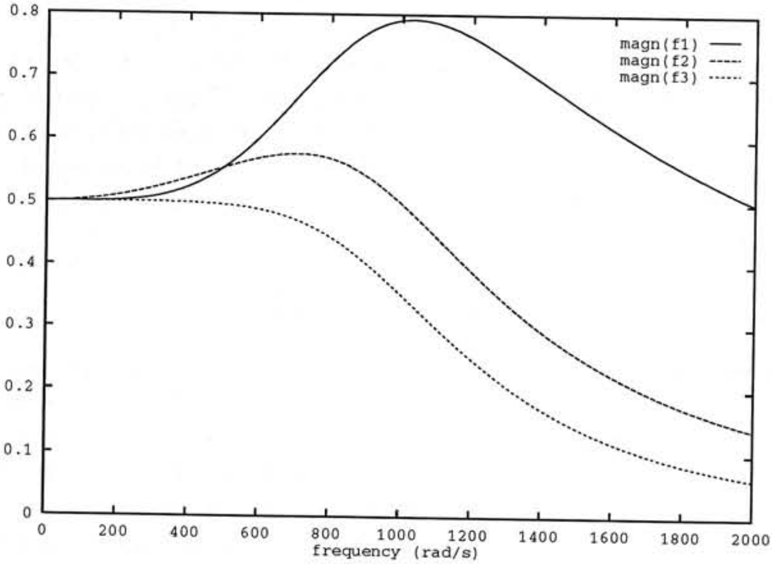


Figure 6.21: Unscaled integrator outputs

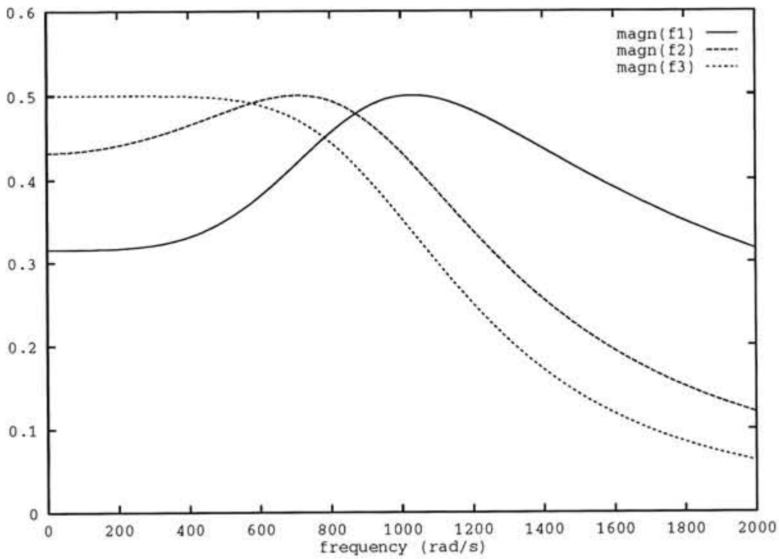


Figure 6.22: Scaling to the tops

Another possibility of scaling is that of scaling on the integral over the signal outputs over the total frequency spectrum. This can be viewed as the total signal level the integrator should be able to handle, which is represented by the total area under the output level of a single integrator. Such scaling results in the frequency characteristic shown in fig.6.23. This optimization is most suitable for white noise input signals, and is, for example, a good model for a radio input spectrum.

The optimization carried out only considers the output signal handling capabilities of the integrators. This is usually called scaling. Referring to the state space description, the transfer from the input of the filter to the output of the integrators is equal to:

$$F = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} = (sI - A)^{-1}B \tag{6.29}$$

From this the "controllability matrix" can be constructed:

$$K = \frac{1}{2\pi}\int_{-\infty}^{\infty} FF^* d\omega \tag{6.30}$$

This equation can also be obtained by the recursive formula:

$$AK + KA^T = -BB^T \tag{6.31}$$

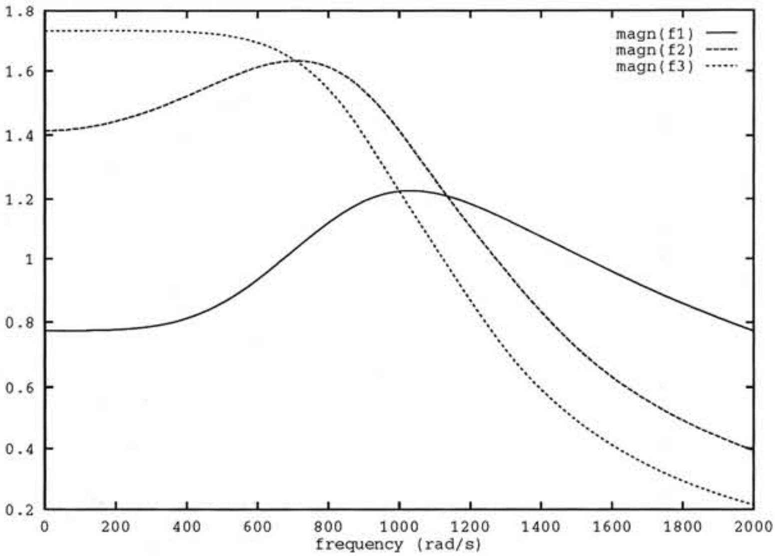


Figure 6.23: Scaling to the power transfer

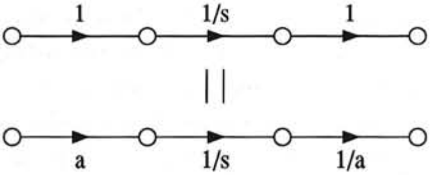


Figure 6.24: Effect of scaling

This matrix describes the (power-)transfer from the input of the filter to the output of the integrators.

Scaling does not have any effect on the topology of the filter. This is clear as the scaling action is as shown in fig.6.24. By equalizing the output levels of all integrators, no single integrator will limit the dynamic range at the upper limit. This is the basic idea behind scaling. It is clear that scaling actually replaces a branch in the filter by another branch with another amplification factor. By means of this amplification factor, the actual integrator is optimally used, with respect to output capability, without changing the topology.

Full optimization –in contrast to scaling– makes use of the noise transfer of the integrators to the output too. Not only is the output capability of the integrators important, but also the noise level. Equalizing the noise levels as well as keeping the output signal capabilities equal makes full optimization feasible. The transfer of the noise sources at the inputs of the integrators to the output of the filter is described as:

$$G = (g_1 \cdots g_n) = C(sI - A)^{-1} \tag{6.32}$$

Now the observability matrix W can be defined as:

$$W = \frac{1}{2\pi} \int_{-\infty}^{\infty} GG^* d\omega \tag{6.33}$$

This matrix can be found recursively too by:

$$A^T W + WA = -C^T C \tag{6.34}$$

To minimize the total noise of the filter, the main diagonal of the matrix W must be equalized. The main diagonal entries can be viewed as the noise transfers of the inputs of the various integrators to the output of the filter. The difference in the noise of the various integrators also has to be taken into account. This can be done by choosing an ideal capacitance division over the integrators.

Because both the W and K matrices are used to optimize the filter, the topology changes. The result is usually a filter with non-zero coefficients in every matrix entry. Thus, a fully connected network of integrators evolves. This is most often a structure too large and difficult to implement on a chip. The dynamic range is often only used to compare with the actual scaled design, which gives an indication of the quality of the filter structure used. Because a method is known to obtain a topology that actually yields an optimal filter, it is important to know the fundamental limits in advance, so that a filter designer is able to know in advance if it is possible to realize the specifications.

6.7.3 Fundamental limits

It is possible to derive fundamental limits for the dynamic range of bandpass filters. The dynamic range can be maximally:

$$DR_{opt} = \frac{V_{max}^2 C}{4kT\xi Q} \cdot f(H(j\omega)) \tag{6.35}$$

The first part of the expression shows that the output capability and the total capacitance increase the dynamic range when they are enlarged. The noise factor of the active components must be as small as possible. It is also clear that the Q of the filter should be kept as low as possible from dynamic range point of view. The second part of this equation is only dependent on the transfer function of the filter.

The total minimal power consumption is also known. The maximal signal level is equal to V_{max}. Current flows optimally only through the integration capacitors, at a frequency of at most ω_c, the cut-off frequency of the filter. Thus the supply current becomes:

$$I_{sup_i} = \frac{\omega_c C_i V_{max}}{\pi} \tag{6.36}$$

This is the minimal current through integrator i, because the current through the resistors is not taken into account, nor is the surplus current of the biasing for the active circuits.

6.8 Tuning

The function of filters is to separate signals on the basis of frequencies. The frequencies of interest have to be passed through the filter, as other frequencies are attenuated. The specifications are defined by the system in which the filters are used. As the components in circuits can, for example, have a 20 % tolerance, filters have to be tuned. Often, filters have to be tuned over a certain frequency range, for example, in a radio receiver. The above two comments imply that tuning to some reference frequency is desirable. In order to accomplish tuning, a reference frequency has to be available, and the filter must have the possibility of being tuned. The filters described in earlier sections can be tuned in two ways. Firstly, by varying the capacitance in the integrators. This is possible when using junction capacitances, but this causes problems, because junction capacitances have a polarity, they usually cannot be used floating and they are strongly non-linear with voltage. Secondly, the transfers G_x can be used to tune the filters. In practice, this mostly means that resistors are varied. Examples are MOS transistors in triode, which can be tuned by varying the gate voltage.

There are several tuning options, two of which are discussed below. Both methods rely on the matching of components, which determine the time constants of the filter. A Voltage Controlled Oscillator (VCO) or a reference filter can be used for tuning. Both these methods are depicted in fig.6.25.

In both cases, the same components are used in the VCO or the reference filter as in the desired filter. If, for example, an eighth-order bandpass filter has to be tuned, a second-order reference filter can be chosen, with a center frequency equal to the filter to be tuned. It is also possible to use a VCO (undamped filter) that oscillates at the frequency the filter should be tuned at. By means of a feedback loop, the VCO or reference filter can be tuned. Because of the matching to the desired filter of the VCO or the reference filter, the master filter is tuned too. The two possible feedback loops are shown in fig.6.26.

Of the two methods presented, the VCO tuning has the advantage of not being sensitive to phase errors due to, for example, the phase comparator. VCO nonlinearities, however, influence the tuning error. The realized loop actually is a classic phase-locked loop (PLL).

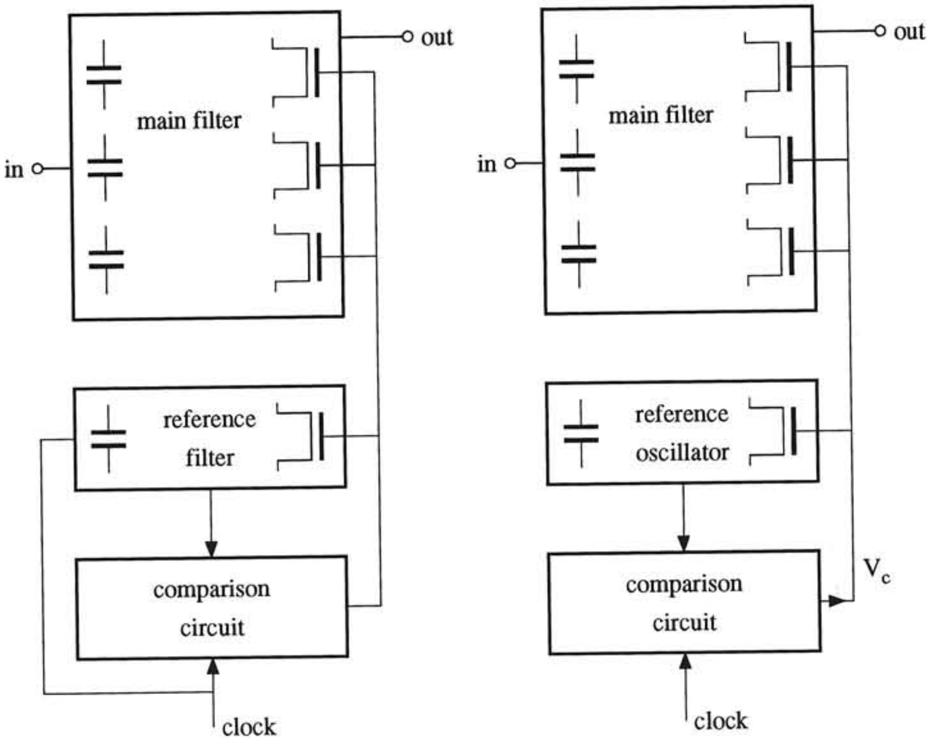


Figure 6.25: Two methods of tuning filters

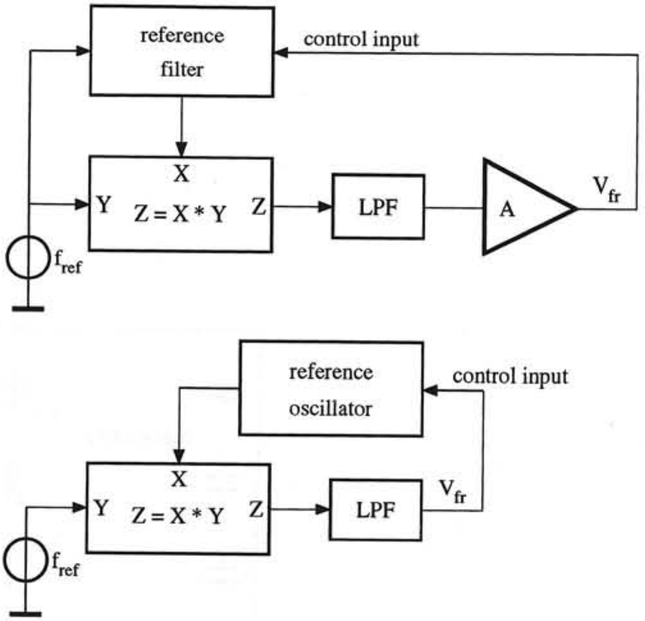


Figure 6.26: Two feedback loops for tuning filters

6.9 Problems

Methods for designing filters all assume that the elements to construct a filter are ideal. In practice, this assumption is not true, and this has to be taken into account by the filter designer. The problems a filter designer encounters are best illustrated by a real integrator design. Take, for example, a transconductance-C integrator. It is desirable that the transconductance has infinite input impedance and infinite output impedance. In addition, ideal integrators have an infinite bandwidth. Having no parasitic capacitances at the input and the output ensures that these high frequencies can be spanned. The choice of the transconductance-C $(g_m - C)$ integrator ensures the parasitic capacitances of the used MOS transistors in parallel with the desired integration capacitance. Suppose, for example, a parasitic drain-source capacitance in the active saturation integrator of figure 6.13. This implies that these capacitances can be used for the transfer of the integrator. The ideal transfer yields:

$$H_i(s) = \frac{V_o(s)}{V_i(s)} = \frac{g_m}{sC} \tag{6.37}$$

with $s = j\omega$. The integrator has infinite dc amplification, and no parasitic poles or zeros. This implies an integrator phase of -90^o over the whole frequency region.

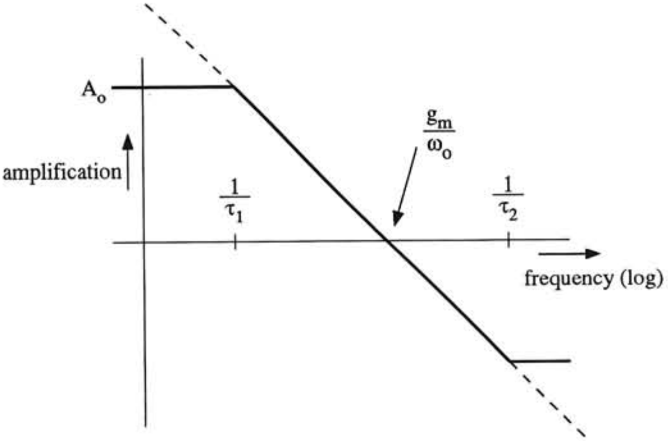


Figure 6.27: Ideal transfer and non ideal transfer

The frequency of unity gain is at:

$$\omega_T = \frac{g_m}{C} \tag{6.38}$$

The transfer of the integrator with parasitic poles and zeros can be described as:

$$H_{ni} = \frac{Vo(s)}{V_i(s)} = A_o \cdot \frac{1 - s\tau_2}{1 + s\tau_1} \tag{6.39}$$

The transfer is shown in fig.6.27.

Although there are more parasitic poles and zeros, only one zero is chosen to represent the phase lead or lag. In order to model the phase error accurately, the zero can be placed in the left or the right half-plane. To accomplish an accurate filter transfer, the phase is not allowed to deviate much from $-90°$, only some tenths of a degree, depending on the transfer function. This is a very important design aspect of the integrators.

6.9.1 Effects of non-idealities

The effects of the non-idealities show themselves in the final filter design. The various non-idealities all have their own influence on the filter transfer. The effect of small dc gain occurs most dominantly in bandpass filters with high Q: it decreases the Q of the filter.

Parasitic poles and zeros influence the position of the poles of the desired transfer. This can result in a Q enhancement, but also a Q decrease. It is even possible that the filter becomes an oscillator. In that case the poles are driven to the right half plane.

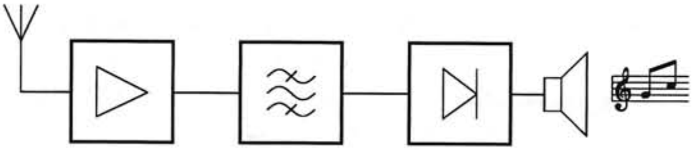


Figure 6.28: Example of direct receiver

Many other non-idealities are possible. The effects of these are often not easy to find in the transfer function. An important effect is non-linearity.

6.10 Example of a receiver with active filters

Until recently, active filters could not be employed in radio receivers, because of their low dynamic range. Fully integrated receivers almost always use external filters for selectivity. Considering fully long-wave receivers (150 kHz to 300 kHz), it appears to be possible to use direct receivers with moderate quality, see fig.6.28. Using a channel bandwidth of 7 kHz at 100 kHz gives a Q of 15. Using filters with a total capacitance of 80 pF and a supply voltage of 3V, a dynamic range of 80 dB is possible.

Attempts to design medium-wave receivers with the same capacitance and supply voltage have had dynamic range problems. As the channel bandwidth is 7 kHz too, and the center frequency is around 1 MHz, the resulting Q is about 10 times as large. This gives a dynamic range deterioration of 20 dB. The power consumption also increases by a factor of 10, because the same integrator capacitances are applied at a frequency 10 times as high.

A high-quality medium-wave receiver requires a dynamic range of at least 120 dB, as is usual in car radio. To realize these specification with active filters, it is necessary to use at least 806 nF total capacitance with a minimal power consumption of 5.2 W. It is currently not possible to realize these specifications.

6.11 Conclusions and considerations

Since the early days of filter history, there has been much progress. After the introductions of the first non-integratable filters, it was soon discovered that active filters could be made. Only after the start of monolithic technology, however, has the use of active filters become widespread. The use of coils is not a very attractive alternative in current monolithic technologies, so special structures have to be used to overcome the associated problems. Active structures limit the dynamic range of the filters. As chip area is an important economic measure, it is important that the total of chip area used is limited. This requires design methods that profit fully from the available components. Quite elaborate mathematics are necessary to calculate these maxima. More problems are encountered in the non-idealities of components. We think of parasitic capacitors or non-linearities. These problems become increasingly important at higher frequencies and lower power supply voltages.

Chapter 7

Automatic gain controls

Wouter Serdijn

7.1 Introduction

Automatic gain controls (AGCs) are widely used in communication systems to modify the dynamic range of a signal. They can be found in, e.g., radio receivers and transmitters, audio amplifiers and hearing instruments.

An AGC is a circuit that automatically controls its gain in such a way that variations in the input signal result in smaller variations in the output signal. This control action is usually performed by means of a loop that contains a large time constant (e.g. several tens of milliseconds).

In the past, this large time constant was realized by means of a large (external) capacitor. However, in integrated circuit implementations, external components should be avoided as much as possible.

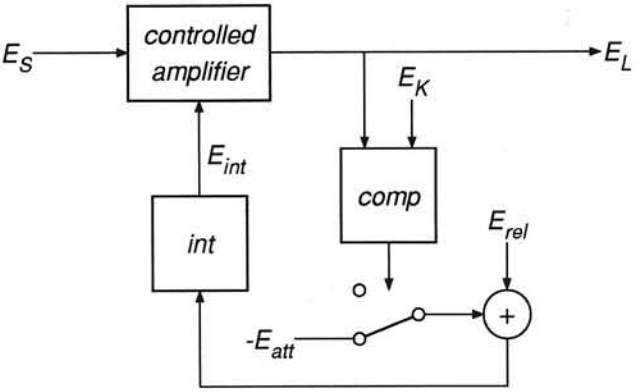


Figure 7.1: Block diagram of an automatic gain control ($C.R. = \infty$)

A typical AGC circuit is shown in Figure 7.1. The output signal E_L is compared with a reference level E_K (the knee level) by a comparator (*comp*) that determines whether the integrating circuit (*int*) — in practice often nothing more than an RC network — is charged (by $E_{\text{att}} - E_{\text{rel}}$) or discharged (by E_{rel}). The output signal of the integrator, E_{int}, forms the control signal of the controlled amplifier. The operation is as follows: If E_{att} is larger than E_{rel}, the output signal E_L is controlled toward the knee level E_K. Variations in the input signal therefore always result in smaller or equal variations in the output signal. See, e.g., Figure 7.2. The control action requires some time. This can be described by the expressions *attack time* and *release time*. The attack time is defined as the time needed for the AGC to respond to a sudden 25 dB increase in the input signal until the output signal is within 2 dB from its final value [1]. Vice versa, the release time is defined as the time required to respond to a sudden 25 dB decrease in the input signal until the output signal is within 2 dB from its final value.

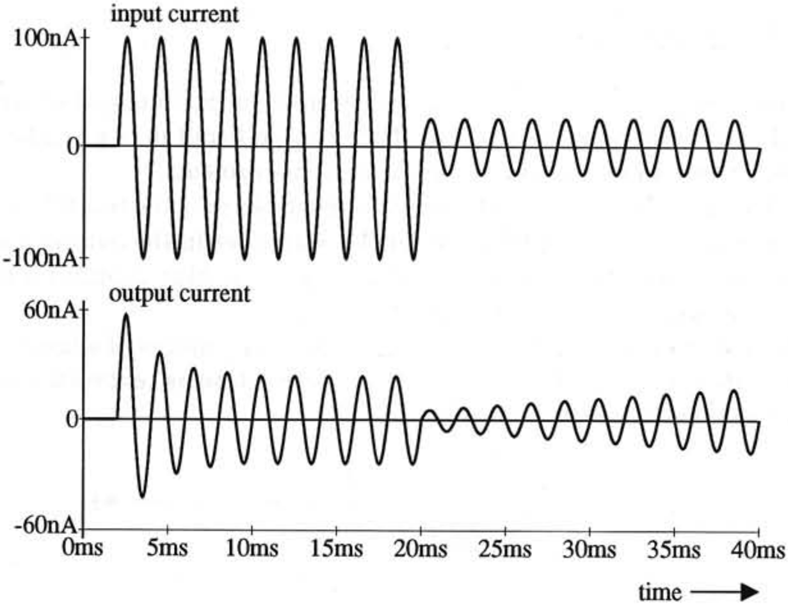


Figure 7.2: Response of an AGC with an infinite compression ratio to a typical input signal

Another important parameter is the *compression ratio* (*C.R.*), defined as the ratio of the variation in the input signal and the variation in the output signal (both in dBs), or

$$C.R. = \frac{\Delta E_{S,\text{dB}}}{\Delta E_{L,\text{dB}}} \tag{7.1}$$

From this expression it can be seen that for $C.R. > 1$ the operation is that of a compressor and for $C.R. < 1$ of an expander.

The circuit given in Figure 7.1 realizes an infinite compression ratio, since $\Delta E_L = 0$ dB. In the following section, realizations with different compression ratios are discussed.

7.2 AGCs with finite compression ratios

For AGCs with finite compression ratios, the output signal E_L cannot directly be compared with the reference level E_K. We thus need at least one additional amplifier to generate an additional signal out of E_L, E_S or E_K.

7.2.1 Controlled amplifiers in cascade

One way of obtaining a finite compression ratio is to pass the output signal of the AGC, E_L, through another controlled amplifier, which is controlled by the same control signal E_{int}. The output of this second amplifier can then be compared with E_K and thus kept constant. This situation is depicted in Figure 7.3.

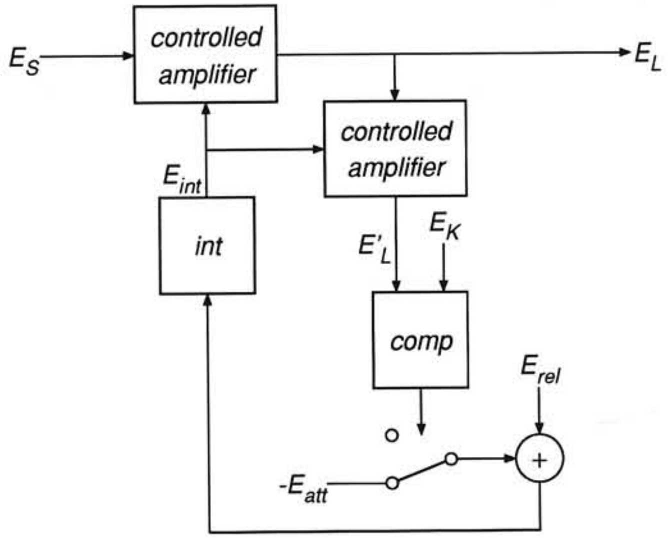


Figure 7.3: AGC with $C.R. = 2$ using two controlled amplifiers in cascade

If the controlled amplifier consists of a simple multiplier three equations can be extracted.

$$E'_L = E_K \tag{7.2}$$
$$E_L = E_S E_{\text{int}} \tag{7.3}$$

$$E'_L = E_L E_{\text{int}} \tag{7.4}$$

with E_L the output signal of the AGC, E'_L the output signal of the second controlled amplifier, which is kept equal to E_K. These equations can be rewritten as follows:

$$E'_L = E_S E_{\text{int}}^2 \tag{7.5}$$

$$E_{\text{int}} = \sqrt{E_K/E_S} \tag{7.6}$$

$$E_L = \sqrt{E_K E_S} \tag{7.7}$$

For the compression ratio of the AGC we thus can write

$$C.R. = \frac{\Delta E_{S,\text{dB}}}{\Delta E_{L,\text{dB}}} = \frac{20 \log E_S}{20 \log \sqrt{E_K E_S}} = 2 \tag{7.8}$$

because E_K is a constant DC level.

Realizing compression ratios other than two is done by using additional controlled amplifiers in cascade. However, because of the greater complexity, this is believed to be of little practical value.

7.2.2 Differently controlled amplifiers

Another possibility is making use of two controlled amplifiers that both have the same input signal E_S, but are controlled by different control signals. This is depicted in Figure 7.4. The output signal of the integrator, E_{int}, is passed to the controlled amplifier that generates the output signal and to a multiplier that multiplies E_{int} by a constant factor m. This multiplied version of E_{int} is then passed to the second controlled amplifier that generates the signal that is to be compared with E_K.

In order for it to operate properly, the input-output relation of the controlled amplifiers cannot be that of a multiplier for this would result in an infinite compression ratio. We therefore assume that both amplifiers realize an exponentially controlled transfer function, or

$$E_{\text{out}} = E_{\text{in}} \exp E_{\text{control}} \tag{7.9}$$

This transfer function can easily be realized, as is shown in Section 7.4. For the circuit shown in Figure 7.4 again three expressions can be found.

$$E'_L = E_K \tag{7.10}$$

$$E_L = E_S \exp E_{\text{int}} \tag{7.11}$$

$$E'_L = E_S \exp mE_{\text{int}} \tag{7.12}$$

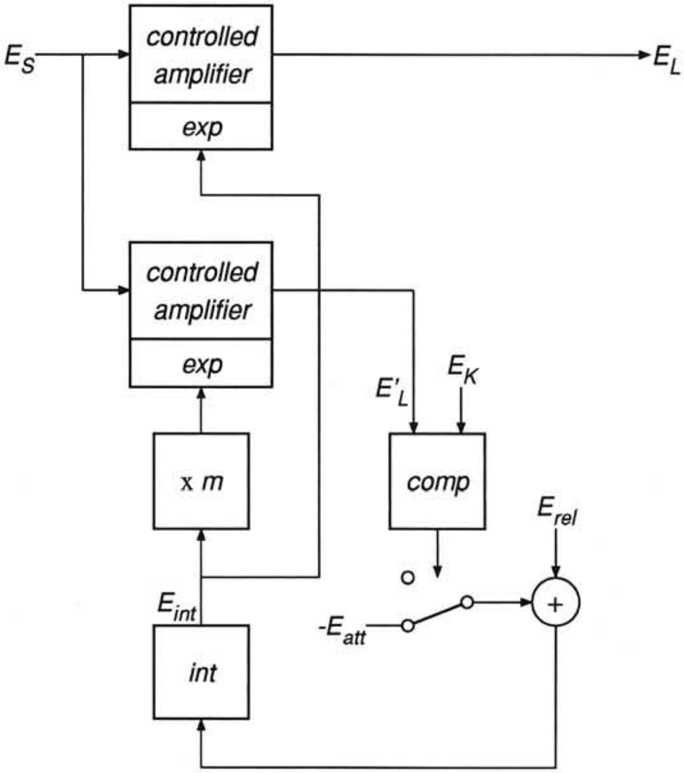


Figure 7.4: AGC with $C.R. = \frac{m}{m-1}$ using differently controlled amplifiers

These can be rewritten as

$$
\begin{aligned}
E'_L &= E_K && (7.13)\\
E_{\text{int}} &= \frac{\ln E_K/E_S}{m} && (7.14)\\
E_L &= E_S \exp \frac{\ln E_K/E_S}{m} && (7.15)\\
&= E_S^{1-1/m} E_K^{1/m} && (7.16)
\end{aligned}
$$

For the compression ratio we then find

$$
C.R. = \frac{\Delta E_{S,\text{dB}}}{\Delta E_{L,\text{dB}}} = \frac{1}{1-1/m} = \frac{m}{m-1} \qquad (7.17)
$$

Using this technique, all compression factors between zero and infinity can be realized. A compression ratio of two, for example, is thus obtained by choosing $m = 2$.

7.2.3 Controlled knee level

Finally there is also the possibility of passing the reference level E_K through another controlled amplifier and comparing its output signal to the output signal of the AGC. This is depicted in Figure 7.5. As E_K contains no signal information (i.e. is a constant DC level) the demands that are made upon the second controlled amplifier can be much less, thereby reducing the circuit complexity.

Again both amplifiers are exponentially controlled. The output signal of the integrator, E_{int}, is passed to the controlled amplifier that generates the output signal of the AGC and to a divider that divides E_{int} by a constant factor m. This divided version of E_{int} is then passed to the second controlled amplifier that generates the signal that is to be compared with the output signal E_L from the reference level E_K. Again three expressions can be found.

$$
\begin{aligned}
E'_K &= E_L && (7.18)\\
E_L &= E_S \exp E_{\text{int}} && (7.19)\\
E'_K &= E_K \exp E_{\text{int}}/m && (7.20)
\end{aligned}
$$

These can be rewritten as

$$
\begin{aligned}
E'_K &= E_L && (7.21)\\
E_{\text{int}} &= m \ln E'_K/E_K && (7.22)
\end{aligned}
$$

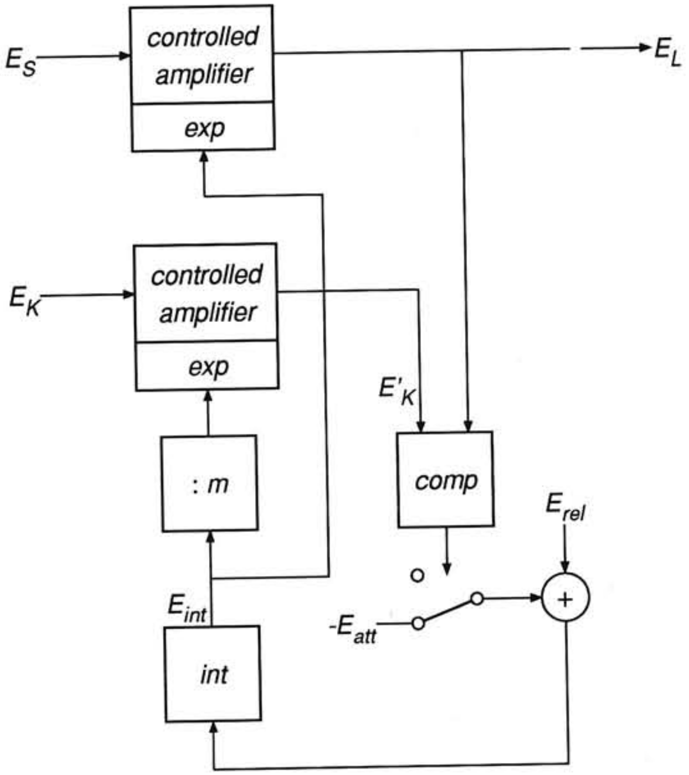


Figure 7.5: AGC with $C.R. = 1 - m$ using a controlled knee level

$$= m \ln E_L/E_K \tag{7.23}$$

$$E_L = \frac{E_S^{1/(1-m)}}{E_K^{m/(1-m)}} \tag{7.24}$$

For the compression ratio this results in

$$C.R. = \frac{\Delta E_{S,\mathrm{dB}}}{\Delta E_{L,\mathrm{dB}}} = \frac{1}{1/(1-m)} = 1 - m \tag{7.25}$$

A compression ratio of two, for example, is obtained by choosing $m = -1$. Hence, in this situation the divider is an inverter.

7.3 AGCs in the current domain

In Chapter 2 is has been shown that low-voltage low-power integrated circuits for preference operate in the current domain. For this reason, current is chosen as the information-carrying quantity of the various subcircuits as much as possible. However, we see in the next section that the exponentially controlled amplifiers proposed here are controlled by means of a voltage. As the only integratable integrating element is a capacitor, and its input signal is a current, whereas its output signal is a voltage, the integrator will consist of a capacitor followed by a voltage follower. This voltage follower generates a low-impedance version of the voltage across the capacitor to prevent interaction between the capacitor and the controlled amplifier.

7.4 Controlled current amplifiers

Controlled amplifiers can be divided into two different types. First there is the class of controlled amplifiers of which the output signal shows no significant variation, but of which the input signal varies over a wide range. As an example, we mention an AGC with infinite compression; its input signal varies significantly but the output signal is almost unchanged. Second, there are controlled amplifiers of which the input signal shows no significant variation, but of which the output signal varies over a wide range. For example, in an ordinary audio amplifier; the output signal is controlled so that the sound pressure level corresponds to the need of the listener.

A well-known and commonly used controlled amplifier is the *differential pair* of which the transconductance is controlled by varying its tail current. However, as the input signal is limited to some tens of millivolts, while the output signal can be made to vary over a wide range by simply adjusting the tail current, it falls into the category of the second type and therefore it is clearly not the best type of controlled amplifier to use in an AGC. Apart from the above disadvantages, its input signal is a voltage which makes the differentail pair less suitable for our purposes.

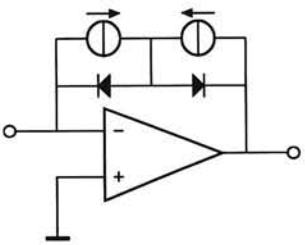


Figure 7.6: Transimpedance amplifier with two diodes in anti-series in the feedback path

Another example of a controlled amplifier is given in Figure 7.6: a transimpedance amplifier of which the feedback network consists of two diodes in anti-series. The transfer function equals the sum of the dynamic resistances of both the diodes, which can be varied by controlling the bias currents through the diodes. Although it is of the first type of controlled amplifiers — the output voltage swing is limited, while the input current swing is not —, the problem with this circuit is that, apart from the bias current, also signal current flows through the diodes, thereby varying the dynamic resistances and distortion occurs. To reduce this distortion, the biasing currents must be much larger than the signal current which degrades the power efficiency. Further, its output signal is a voltage which additionally makes this type of amplifier less suitable for our purposes.

7.4.1 Four fundamental ways of controlling the gain

A suitable solution is a current amplifier of which the gain equals the ratio of two transconductances: the scaling current amplifier (Figure 7.7) [2]. As the transconductance of a bipolar transistor is proportional to its (DC) collector current, we can vary the gain by varying the collector current of either Q_1, I_{C,Q_1}, or Q_2, I_{C,Q_2}. We can say that the controlled amplifier is of the first type if I_{C,Q_1} is controlled (I_{C,Q_2} remains constant, thus limiting the output current swing) and to the second type if I_{C,Q_2} is controlled (I_{C,Q_1} remains constant, thus limiting the input current swing). A theoretical possibility is that both the collector currents are controlled. However, this option is believed to be of little use in practice.

Another way of controlling the ratio of the transconductances, and thus the gain of the amplifier, is by means of a controlling voltage V_C connected between the emitter of Q_1 and the emitter of Q_2. See Figure 7.8. We now obtain a gain A_i that is proportional to the anti-log of V_C, or

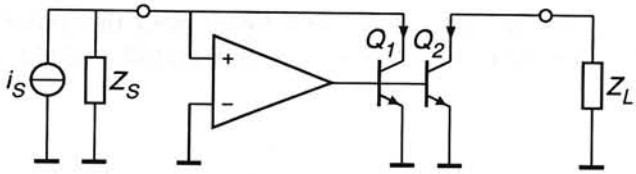


Figure 7.7: Controlling the gain of a scaling current amplifier by controlling the ratio of the collector currents of Q_1 and Q_2

$$A_i = -g_{m,Q_2}/g_{m,Q_1} = -e^{V_C/V_T} \approx 335V_C \text{ dB, } V_C \text{ in volt} \qquad (7.26)$$

in which V_T equals the thermal voltage kT/q, approximately 26 mV at 300 K.

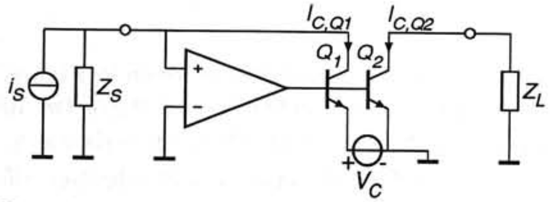


Figure 7.8: Controlling the gain of a scaling current amplifier by controlling voltage source V_C

The exponential relationship between the gain A_i and the control voltage V_C of the voltage-controlled amplifiers enables us to control the gain over a wide range with only a small control voltage.

As the controlled current amplifiers are either current- or voltage-controlled and are of either the first or second type, we can distinguish four different kinds:

- a current-controlled type 1 scaling current amplifier,
- a current-controlled type 2 scaling current amplifier,
- a voltage-controlled type 1 scaling current amplifier, and
- a voltage-controlled type 2 scaling current amplifier.

These four are the subject of the next four subsections. Unless there is the possibility of on-chip filtering, the biasing of a circuit is by preference done by setting the common-mode quantities [2]. In order to do so, the signal path has to be symmetrical. As current- or voltage-controlled and type 1 or type 2 has nothing to do with the signal behavior of the amplifier, we assume that the design of the symmetrical signal path has been completed in an earlier stage and start our considerations from here.

It also is tacitly assumed that the source is floating and the load is tied to a certain reference level V_Z (e.g. a base-emitter voltage of the following circuit). In other situations, similar solutions can be found.

7.4.2 The current-controlled type 1 symmetrical scaling current amplifier

The general biasing solution for a current-controlled type 1 symmetrical scaling current amplifier is depicted in Figure 7.9. The transfer function is controlled by means of two current sources I_C. In order to make the DC collector currents of the output transistors equal to I, a common-mode output is generated by two extra output transistors. The sum of their collector currents is compared with $2I$, thus producing an error signal. The error signal is amplified by the op amp and fed back to both emitters of the input transistors, thereby setting the correct emitter current. As the absolute value of the loop gain of the common-mode loop is much larger than one, the error signal is nullified and the output transistors are biased correctly.

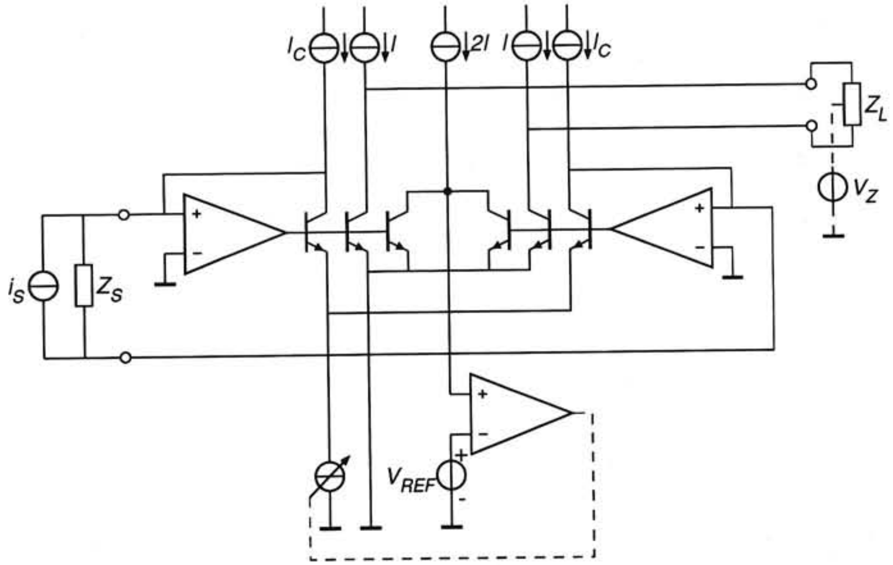


Figure 7.9: Current-controlled type 1 symmetrical scaling current amplifier

7.4.3 The current-controlled type 2 symmetrical scaling current amplifier

This situation does not differ much from the preceding one. Only now the transfer function is controlled by varying the current through the *output transistors.* The general solution for a current-controlled type 2 symmetrical scaling current amplifier is depicted in Figure 7.10. In order to set the output collector currents, again a common-mode output is generated by two extra transistors, their collectors tied together. The common-mode current is compared with $2I_C$, producing an error signal. The error signal is amplified by the op amp and fed back to both emitters of the *output transistors*, thereby setting the correct emitter current. As the absolute value of the loop gain of the common-mode loop is much larger than one, the error signal is nullified and the output transistors are biased correctly.

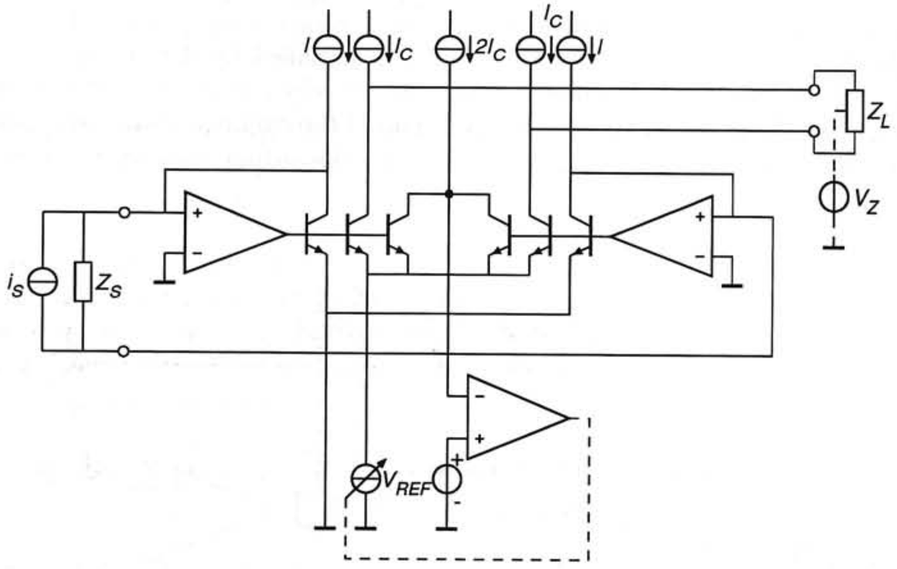


Figure 7.10: Current-controlled type 2 symmetrical scaling current amplifier

7.4.4 The voltage-controlled type 1 symmetrical scaling current amplifier

The transfer function of this type of amplifier is controlled by the control voltage V_C (see Figure 7.11). Again a common-mode replica of the common-mode collector currents through the output transistors is compared with $2I$, producing an error signal. The error signal is amplified by the op amp and controls the collector current of the input stages.

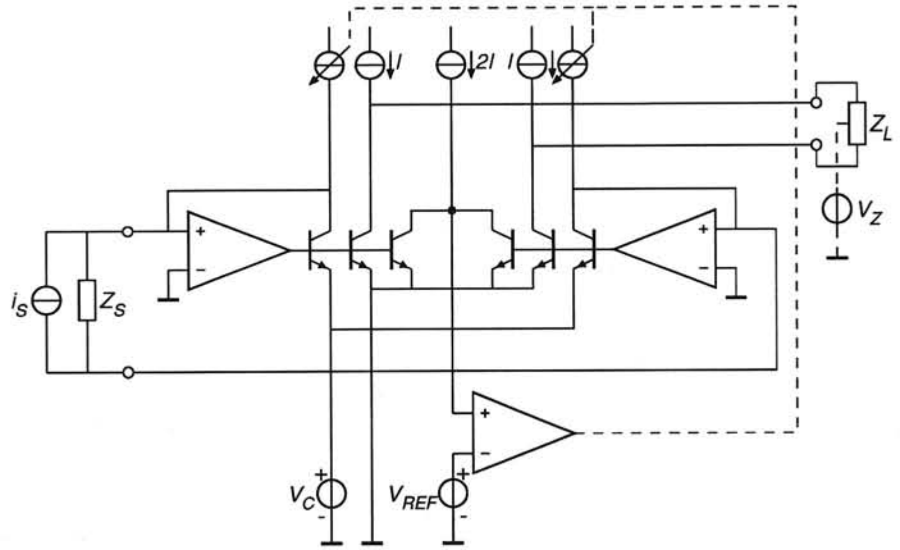


Figure 7.11: Voltage-controlled type 1 symmetrical scaling current amplifier

An example of the voltage-controlled type 1 symmetrical scaling current amplifier is described in [3]. This circuit is part of a hearing instrument and serves to attenuate the signal coming from the preamplifier [4], that can vary between 40 nA and 10 μA (peak value), and drive a highpass filter [5] at maximal 25 nA.

7.4.5 The voltage-controlled type 2 symmetrical scaling current amplifier

Finally, the voltage-controlled type 2 symmetrical scaling current amplifier, see Figure 7.12. V_C controls the gain. The common-mode loop controls the collector currents of the output transistors (including the common-mode output transistors) in such a way that they match the collector currents of the input stages.

Two examples of the voltage-controlled type 2 symmetrical scaling current amplifier can be found in [6] and [7]. The first circuit (though not strictly symmetrical) has been designed for the same hearing instrument and loads a highpass filter [5] at maximal 25 nA and drives the power amplifier. Its gain can be controlled from 0 to 60 dB. The second circuit, a controllable preamplifier, was originally designed for a different hearing instrument using the conventional electret microphone with built-in JFET [8]. It adapts input signals varying between 120 nA and 30 μA (peak value) to the maximal filter input level of 1 μA.

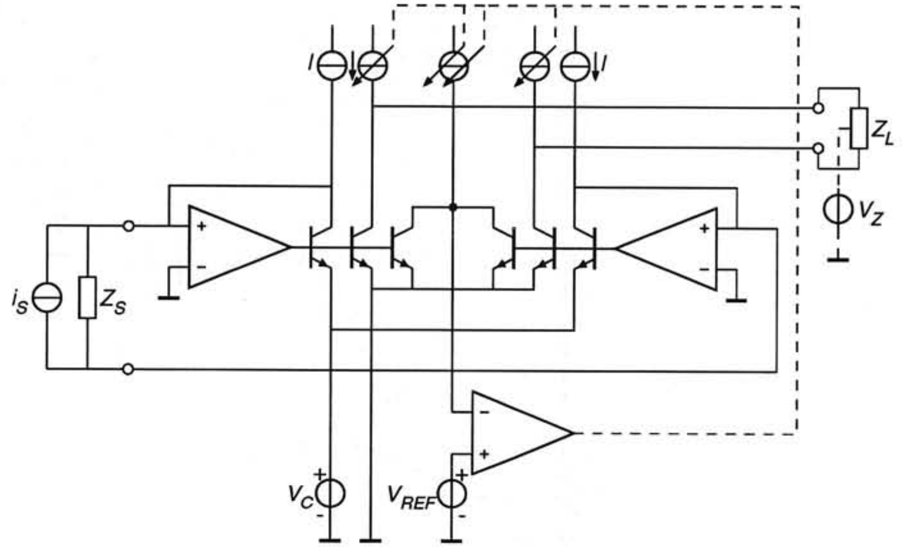


Figure 7.12: Voltage-controlled type 2 symmetrical scaling current amplifier

7.5 Comparators

The comparator is the circuit that compares the output current of the controlled amplifier with the reference level I_K by means of a highly non-linear input-output relation. Thus the comparator can be viewed as a one-bit A/D converter. Its response to a signal level higher than the reference level is a fixed output level, representing the '0' or the '1' state. Its response to a signal lower than the reference level is another fixed output level, representing the complementary state. The gain when the input equals the reference level ideally is infinite. However, it is no use making the gain much larger than the ratio of the desired output swing and the smallest input swing.

For a comparator which has a current-current input-output relation we can choose either a cascade connection of a non-linear one-port and a linear two-port, or an amplifier with a saturating input-output relation.

7.5.1 Cascade of a non-linear one-port and a linear two-port

Examples of (bipolar) non-linear one-ports are diodes and pinch resistors. An example of a (current) comparator consisting of a cascade of a non-linear one-port and a linear two-port is given in Figure 7.13. Both the output current of the controlled amplifier and the reference current can be supplied, with opposite signs, to the same input, and thus subtracted from eachother. If the result is positive, diode D_1 will conduct. The resulting anode-cathode voltage will vary only slightly with respect to the current and thus represent one state of the comparator. If the result is negative, diode D_2 will conduct, resulting in a complementary voltage, representing the complementary state. The resistor R transforms the voltage into a current that is sensed by the (low-impedance input of the) next circuit. Although relatively simple, this comparator is supposed to be of less practical importance in low-power integrated circuits. If, for example, the output current of the comparator is to be as small as 25 nA, and the diode voltages are about .5 V, R must equal 20 MΩ. This value is not easily realized in an integrated circuit.

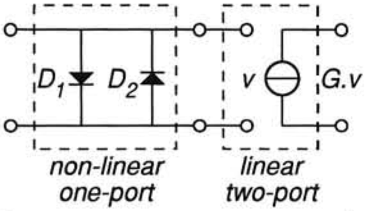


Figure 7.13: Example of a comparator consisting of a cascade connection of a non-linear one-port and a linear two-port

7.5.2 Amplifiers with a saturated input-output relation

It is not difficult to design an amplifier with a saturated input-output relation. Every practical amplifier will come into saturation if its input signal exceeds a certain level. Three examples are given in Figure 7.14. During saturation, often one or more transistors will be pinched off or be in saturation. In negative-feedback amplifiers, this might introduce extra dominant poles in the feedback loop, giving rise to instability, or lead to latch-up (see e.g. [9]). The designer must be aware of this and try to avoid both instability and latch-up under all circumstances. Often good results can be obtained from simple circuitry.

7.6 Voltage followers

The last stage in the design of low-voltage low-power automatic gain controls is the design of the voltage follower. The voltage follower forms a buffer between the capacitance C and the controlled amplifier. Since the input current of a field effect transistor (JFET or MOST) is far below the other currents that charge and discharge the integrating capacitor, these devices are very well suited for this task. When using bipolar transistors this can only be achieved by using negative-feedback techniques. The basic voltage-follower configuration and three possible

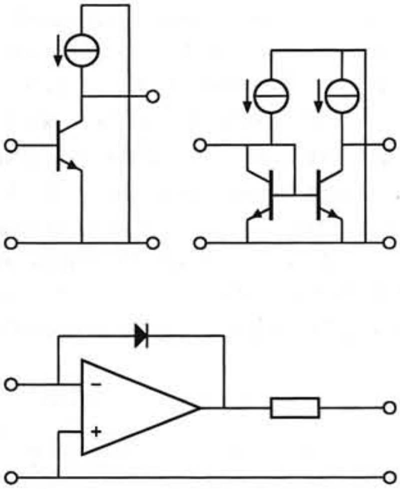


Figure 7.14: Three possible implementations of amplifiers with a saturated input-output relation

implementations, with either one, two or three transistors, are given in Figure 7.15.

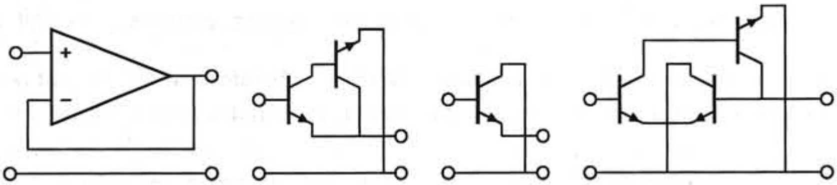


Figure 7.15: Basic voltage-follower configuration and three possible implementations

An example of a three-transistor voltage follower can be found in [3].

7.7 An example: an automatic gain control for hearing instruments

In this section, the design and realization of a low-voltage low-power fully-integratable automatic gain control for hearing instruments is presented.

Apart from pitch, loudness and timbre, information in the world of sound is characterized also by more or less sudden temporary changes. For someone with hearing impairment, these variations do not fit into his or her dynamic range and, therefore, there is either the lack of certain parts of the information or the pain limit is frequently exceeded. In this situation, an automatic gain control can offer certain improvement of the (speech) intelligibility. It must be noted that AGCs are only technically approximate solutions to the dynamic range problems of the hearing impaired. This also explains why optimal, generally applicable values for the AGC characteristics are not easily found.

In practice, two kinds of AGCs are found [10]: an AGC-I and an AGC-O. The AGC-I obtains its control signal from a signal in front of the volume control. Thus, the control action depends on the sound pressure level at the input of the hearing instrument. The control signal of the AGC-O is derived from the signal behind the volume control. The control action then depends on the sound pressure level that is offered to the ear.

The AGCs differ from each other insofar as that the control range of the AGC-I is always larger (for example 60 dB) while for the AGC-O often 20 dB is enough. The compression ratio often is also different: e.g. between 1.5 and 10 for the AGC-I and infinite for the AGC-O.

The circuit that is described here is an AGC-O with an infinite compression ratio, of which the block diagram is shown in Figure 7.1. Its *current-domain* realization is given in Figure 7.16. Apart from the integrator signal E_{int} all signals are represented by currents. The AGC amplifier obtains its input signal from a controllable attenuator [3] and drives a highpass filter [5]. The nominal signal level at both input and output amounts to 25 nA (peak value). The purpose of the AGC is to attenuate larger input signals to avoid clipping and hearing damage. The attack time and the release time must be < 5 ms and 50 ms, respectively. These values are commonly used for hearing instruments.

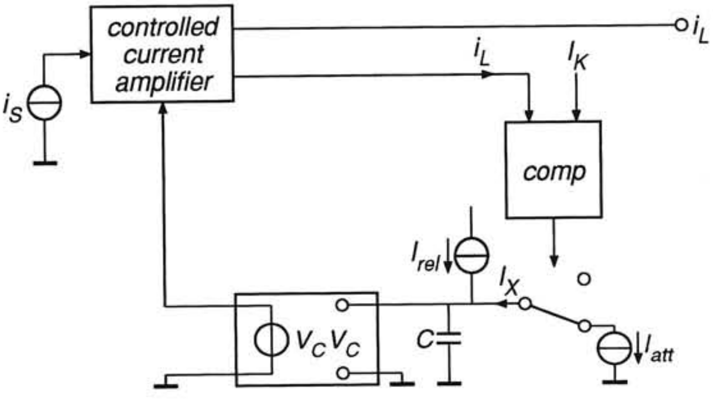


Figure 7.16: Block diagram of an AGC-O operating in the current domain

To here the design has been discussed at system level. We now take a closer look at the design of its components: the controlled amplifier, the comparator (including the switch and current source I_{att}) and the voltage follower.

7.7.1 Design of the controlled amplifier

From the considerations presented in Section 7.4, it should be clear that the best choice of amplifier is a type 1 symmetrical scaling current amplifier. From the two variants, we choose the voltage-controlled one as this gives a control action in dBs, which is perceptibly the most comfortable. A possible implementation of a voltage-controlled symmetrical scaling current amplifier is given in Figure 7.17. As the absolute value of the (differential) loop gain (Q_{1a} and Q_{1b}) is always larger than $B_F/4$, there is no need for additional loop gain; the op amps thus can be replaced by short circuits. We call this a voltage-controlled type 1 symmetrical current mirror. Transistors Q_{1a} and Q_{1b} are the input transistors. Q_{2a} and Q_{2b} deliver the output current i_L. Q'_{2a} and Q'_{2b} are the output transistors for i'_L. The common-mode loop is formed by Q_{3a}, Q_{3b}, Q_4, Q_5, Q_{6a} and Q_{6b}. The collector currents of Q_{3a} and Q_{3b}, which equal the collector currents of Q_{2a} and Q_{2b}, are added and compared with a current $2I$. The error signal controls via Q_4, Q_5, Q_{6a} and Q_{6b} the collector currents of Q_{1a} and Q_{1b}. Because the gain in this loop, the common-mode loop gain, equals the current gain factor B_F of Q_{6a} and Q_{6b}, which is much larger than one, the error signal is nullified and the symmetrical current mirror is biased correctly. Q_{sa} and Q_{sb} limit the maximum gain of the amplifier to one. Q_{da} and Q_{db} shunt the input and prevent the amplifier from saturating.

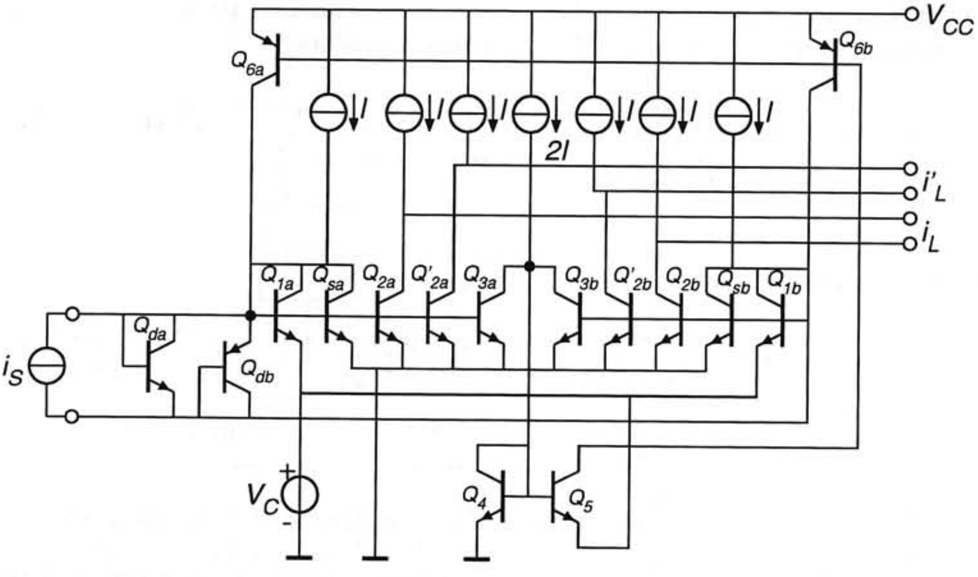


Figure 7.17: The voltage-controlled type 1 symmetrical current mirror

7.7.2 Design of the comparator

The comparator is the subcircuit that decides whether the output current i'_L of the controlled amplifier is larger or smaller than the reference level I_K. For this purpose we can again use a symmetrical current mirror now acting as an amplifier with a saturated input-output relation. Its implementation is given in Figure 7.18. Q_{3a}, Q_{3b}, Q_4, Q_5, Q_{6a}, Q_{6b} and Q_{6c} form the common-mode biasing circuitry. In this case, the common-mode loop gain is kept sufficiently small (i.e. equals 2) to prevent instability in the comparator. The output current I_X therefore switches between 0 and $\frac{2}{3}I$. The two diode-connected transistors Q_{da} and Q_{db} prevent the output transistors Q_{2a} and Q_{2b} from saturating; the comparator switches faster.

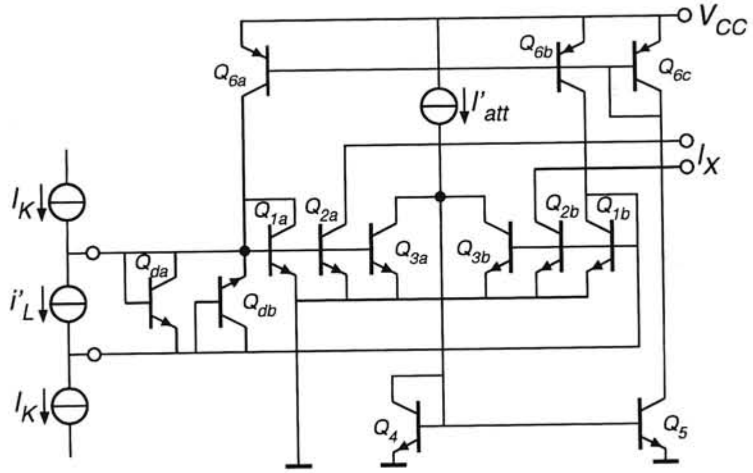


Figure 7.18: A type 1 symmetrical current mirror used as a comparator. The common-mode loop gain equals 2 to prevent instability

7.7.3 Design of the voltage follower

The chosen voltage follower is depicted in Figure 7.19. The input (offset) current equals $I_V/B_{\mathrm{F,PNP}}B_{\mathrm{F,NPN}}$ and must lie well below the release current I_{rel}. Its influence will then be small.

7.7.4 Overall design

Now that all the different parts of the AGC have been designed at circuit level, they can be linked together and we take a look at the numerical values and the remaining bias circuitry.

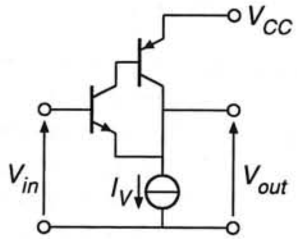


Figure 7.19: The two-transistor voltage follower

As the output signal is to be maximally 25 nA (peak value), the reference current I_K equals 25 nA. The current sources I as depicted in Figure 7.17 have been chosen well above this 25 nA and equal 100 nA. The values of I_{att} and I_{rel} can be derived from the attack time and the release time. Some calculation yields

$$I_{\text{att}} = \frac{5.2V_T C}{t_{\text{att}}} + I_{\text{rel}} \tag{7.27}$$

and

$$I_{\text{rel}} = \frac{2.6V_T C}{t_{\text{rel}}} \tag{7.28}$$

For I'_{att} (Figure 7.18) it follows

$$I'_{\text{att}} = \frac{3}{2}I_{\text{att}} = \frac{7.8V_T C}{t_{\text{att}}} + \frac{3}{2}I_{\text{rel}} \tag{7.29}$$

With t_{att}, t_{rel} and C equal to 4 ms, 50 ms and 400 pF, respectively, this results in 20 nA and 540 pA for I'_{att} and I_{rel}. The current source I_V (Figure 7.19) supplies the current of the PNP transistor in the voltage follower and is chosen to be equal to 1 μA. All these currents can be derived by means of current mirrors with multiple outputs and convenient scaling factors. The scaling factor can be obtained by choosing either a proper emitter area ratio or by means of resistors. The latter solution yields either a *Widlar mirror* or a *gm-compensated mirror* [11].

The total circuit diagram of the AGC is depicted in Figure 7.20. Two voltage sources (V_1 and V_2) have been added to prevent the current sources I_V and I_K from saturating. V_1 has been realized by means of a saturating NPN transistor and a resistor. V_2 contains two saturating PNP transistors in series. Thus, their voltages are well above the saturation voltages of I_V and I_K. To avoid (common-mode) instability, an integratable capacitance C_{comp} can be added.

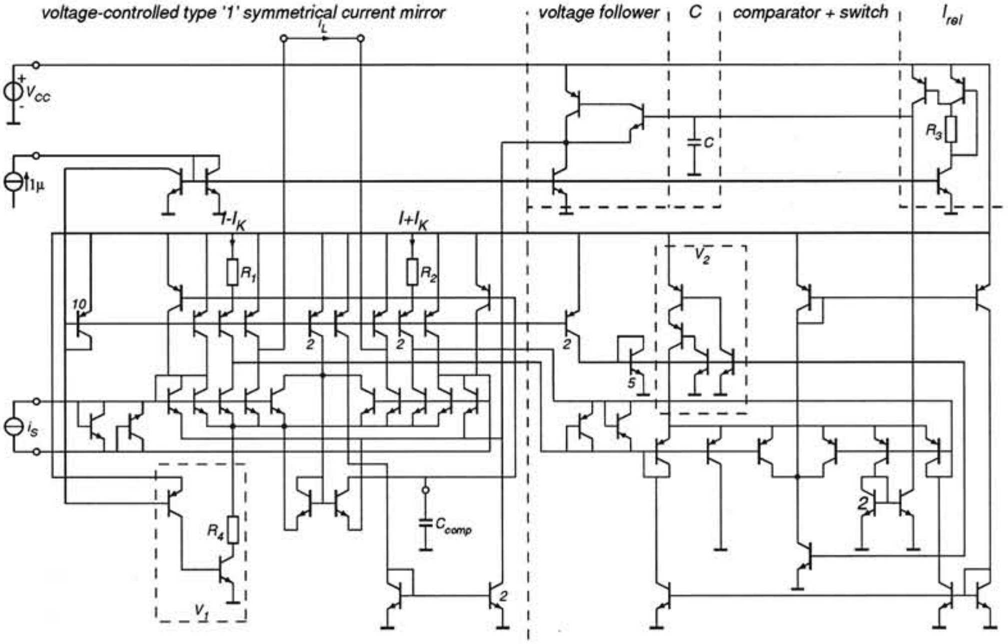


Figure 7.20: The total automatic gain control. Instability may be counteracted by $C_{\mathbf{comp}}$

7.7.5 Experiment results

The active circuitry of the circuit shown in Figure 7.20 has been integrated in the DIMES01 process [12], fabricated at the Delft Institute of Microelectronics and Submicron Technology. Experiments proved the correct operation of the AGC. Table 1 gives the measurement results. No instability occurred. The relatively large value of the release time is caused by the base current of the first stage of the voltage follower. However, this did not pose a problem in our application.

Table 1: Measurement results of the AGC

Parameter	Value	Unit
Compression range	38	dB
Attack time, i_s=1 μA_p, 1 kHz	4.2	ms
Release time, i_s=10 nA_p, 1 kHz	58	ms
Dynamic Range, G=1, B=10 kHz	62	dB
Bandwidth	>100	kHz
Min. supply voltage	1	V
Supply current, G=1	4	μA

References

[1] I.E.C. Recommendation 118-2: *Hearing aids with automatic gain control circuits*, 1983

[2] W.A. Serdijn: *The design of low-voltage low-power analog integrated circuits and their applications in hearing instruments*, Ph.D. Thesis, Delft University of Technology, Delft, the Netherlands, 1994, ISBN 90-6275-955-6.

[3] A. van Staveren: *Low-voltage low-power controlled attenuator for hearing aids*, Electronics Letters, 22nd July 1993, Vol. 29, pp. 1355-1356.

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

[5] W.A. Serdijn: *A low-voltage low-power current-mode highpass leapfrog filter*, Analog Integrated Circuits and Signal Processing 3, 105-112 (1993).

[6] R. Otte and A.H.M. van Roermund : *Low-voltage low-power, wide-range controllable current amplifier for hearing aids*, Electronics Letters, 3rd February 1994, Vol. 30, No. 3, pp. 178-180.

[7] A.C. van der Woerd and W.A. Serdijn: *Low-voltage low-power controllable preamplifier for electret microphones*, IEEE J. of Solid-State Circuits, Vol. 28, pp. 1052-1055, October 1993.

[8] A.G.H. van der Donk: *A silicon condenser microphone: modelling and electronic circuitry*, Ph.D. thesis, Twente University of Technology, Enschede, 1992.

[9] A.G. van Lienden, G.C.M. Meijer and J. van Drecht: *Latch-up in bipolar low-voltage current sources*, IEEE Journal of Solid-State Circuits, Vol. SC-22, pp. 1139-1143, December 1987.

[10] V.J. Geers, F. Keller, A. Löwe and P. Plath: *Technische Hilfe bei der Rehabilitation Hörgeschädigter*, (in German) Springer-Verlag, Heidelberg, 1980.

[11] B. Gilbert : *Bipolar current mirrors*, Chapter 6 in: C. Toumazou, F.J. Lidgey and D.G. Haigh (editors): *Analogue IC design: the current-mode approach*, Peter Peregrinus Ltd., London, 1990.

[12] 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, Vol. 36, pp. 139-147, 1993.

Acknowledgements

It would not have been possible to bring this book to a satisfying close without the help of many people in our professional and personal environment. There are, however, a few people we would like to thank in particular.

Rob Janse prepared many of the drawings in this book, using his self-developed drawing program "tekplot". Jane Zaat-Jones corrected the numerous linguistic errors. Olfien Lefèbre and Yvonne Hom provided helpful secretarial support. David van Maaren and Piet Bos assisted in getting things done at the right moment. Finally, there are the people of the Delft University Press, who allowed a very quick printing schedule. Without their support, this book would not have become what it is. To all we are greatly indebted.

Delft
October 1995

Wouter Serdijn
Chris Verhoeven
Arthur van Roermund

The Authors

Koert van der Lingen was born in De Lier, the Netherlands, on March 11, 1968. He started his course at the Faculty of Electrical Engineering at the Delft University of Technology in 1986 and received his M.Sc. in 1991.

He then joined the Electronics Research Laboratory of the same university, where he has been working towards his Ph.D. His research project aims at understanding and modelling the temperature behaviour of bipolar transistors with respect to their use in bandgap references and temperature transducers.

G.L.E. (Bert) Monna was born in Schiedam, the Netherlands, on September 27th, 1969. He received his M.Sc. in electrical engineering from the Delft University of Technology in 1991.

Since 1991 he has been working towards his Ph.D. on design systematics and continuous-time filters at the Electronics Research Laboratory.

Arthur H.M. van Roermund was born in Delft, the Netherlands in 1951. He received his M.Sc. in Electrical Engineering in 1975 from the Delft University of Technology and his Ph.D. in Applied Sciences from the K.U.Leuven, Belgium, in 1987.

From 1975 to 1992 he was with the Philips Research Laboratories in Eindhoven. First he joined the Consumer Electronics Group, where he was involved with the design and integration of analog circuits and systems, especially switched-capacitor circuits. In 1987 he joined the Visual Communications Group where he has been engaged in video architectures and digital video signal processing. From 1987 to 1990 he was project leader of the Video Signal Processor project and from 1990 to 1992 of a Multi-Window Television project. Since 1992 he has been a full professor at the Faculty of Electrical Engineering of the Delft University of Technology. He heads the Electronics Laboratory, which is part of DIMES: the Delft Institute of Micro Electronics and Submicron Technology.

Wouter A. Serdijn was born in Zoetermeer, the Netherlands, in 1966. He started his course at the Faculty of Electrical Engineering at the Delft University of Technology in 1984, and received his 'ingenieurs' (M.Sc.) degree in 1989.

Subsequently, he joined the Electronics Research Laboratory of the same university where he received his Ph.D. in 1994. His research includes developing a formal design theory for low-voltage and ultra-low-power analog integrated circuits along with the development of circuits for hearing instruments. He has edited two special issues on low-voltage low-power analog integrated circuits for Kluwer's Analog Integrated Circuits and Signal Processing. He teaches electronic design techniques.

Arie van Staveren was born in Hazerswoude, the Netherlands, on April 19, 1968. He received his M.Sc. in electrical engineering from the Delft University of Technology in 1992. Since 1992 he has been working towards his Ph.D. thesis on design systematics at the Electronics Research Laboratory.

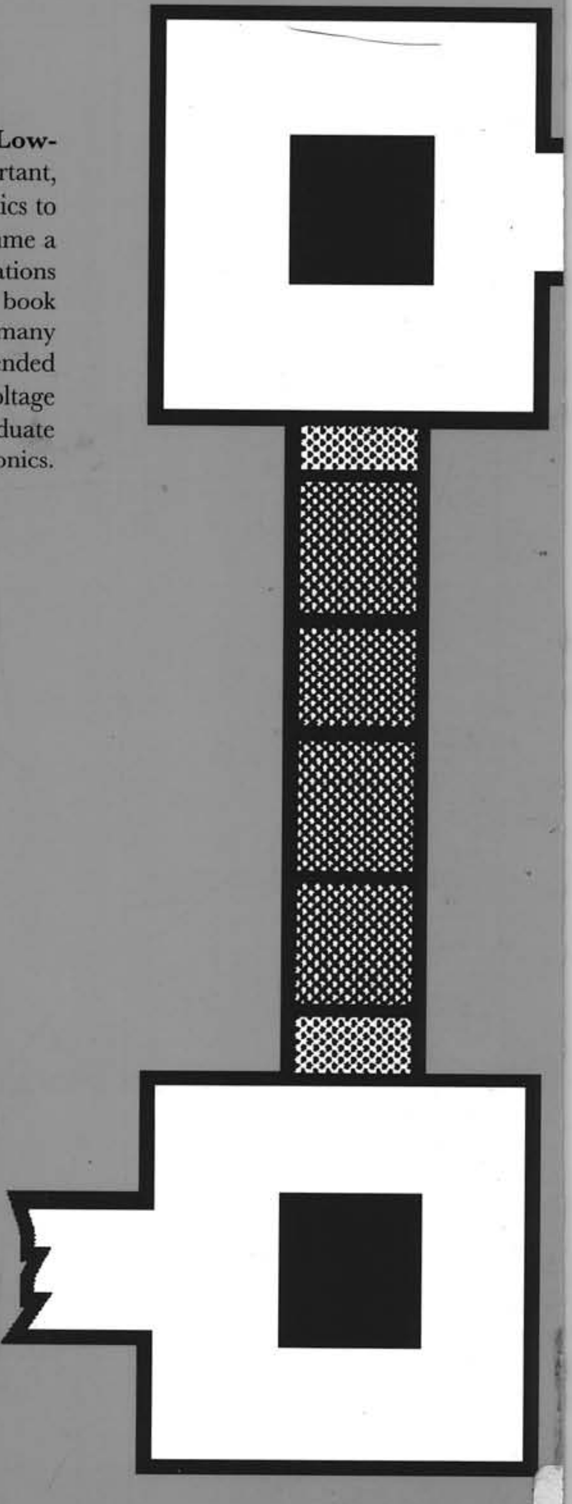
Chris J.M. Verhoeven was born in the Hague, the Netherlands, on February 25, 1959. He received his M.Sc. in electrical engineering from the Delft University of Technology in 1985. In 1985 he joined the Electronics Research Laboratory of the same department in order to prepare a Ph.D. dissertation on "first-order oscillators". He received his Ph.D. in 1990.

At present he is project leader of the group "Structured Electronic Design", at the Electronics Research Laboratory in which the synthesis of analog basic circuits is adressed. To date, the main topics are amplifiers, continuous time filters, bandgap references, oscillators and neural devices.

Albert C. van der Woerd was born in 1937 in Leiden, the Netherlands. In 1977 he received his 'ingenieurs' (M.Sc.) degree in electrical engineering from the Delft University of Technology, Delft, the Netherlands. He was awarded his Ph.D. in 1985.

From 1959 to 1966 he was engaged in research on and the development of radar and TV circuits at several industrial laboratories. In 1966 he joined the Electronics Research Laboratory of the Faculty of Electrical Engineering of the Delft University of Technology. For the first 11 years he carried out research on electronic musical instruments. For the next 8 years his main research subject was carrier domain devices. More recently he has specialized in the field of low-voltage low-power analog circuits and systems. He teaches electronic design techniques.

Analog IC Techniques for Low-Voltage Low-Power Electronics addresses many very important, but recent, techniques which enable electronics to operate at a low supply voltage and consume a minimum amount of power. Apart from investigations at the device, circuit and system levels, the book provides a wealth of practical implementations, many worked out in silicon realizations. The book is intended for both the professional designer of low-voltage low-power analog integrated circuits and the graduate student in this specialized branch of electronics.



ISBN 90-407-1202-6

9 789040 712029

Delft University Press