Vrije Universiteit Brussel

Faculty of Engineering
Department ELEC: Fundamental Electricity and Instrumentation

Design and Analysis of an On-chip Programmable Op-amp Based Filter

Thesis submitted in fulfillment of the requirements for the degree of Master of Science in Applied Sciences and Engineering: Electronics and IT-Engineering (Master of Science in de Ingenieurswetenschappen: Elektronica en informatietechnologie) by

Adam Cooman

Promotor: Prof. Dr. ir. Gerd Vandersteen

2011 – 2012

Dankwoord

Eerst zou ik Gerd willen bedanken voor de deskundige begeleiding van de thesis, ik had geen betere begeleider kunnen indenken. Daarnaast ook bedankt aan Yves en Egon voor de hulp, commentaar en verbeteringen. Zeker ook bedankt aan de hele dienst ELEC voor het creëren van de gezellige sfeer, de liters koffie en de goeie werkomgeving. Ik ben er trots op dat ik mijn thesis bij jullie heb kunnen maken.

Ten tweede bedankt aan iedereen van de familie voor de steun. Ik vind het spijtig dat jullie niets van deze thesis snappen en het onderwerp blijven vergeten, maar misschien slaag ik er ooit in om het duidelijk te maken. Bedankt mama en papa voor alles en ook bedankt aan Io gewoon omdat ze een coole zus is.

Tenslotte, maar zeker niet het minst, zou ik mijn klasgenoten Hannes, Matthias, Zjef, Alexander, Iterio en Pieter willen bedanken. Ik heb mij hard geamuseerd samen met jullie de laatste twee jaar.

"Een tekst is nooit af"

R. Pintelon

Contents

Nomenclature

A	State matrix
B	Input matrix
C	output matrix
D	Feedthrough matrix
μ_0	Charge carrier mobility
C_{ox}	Gate oxide capacitance
f_T	Transition frequency. For MOSFETs, it is defined as the frequency at which theoutput current through a short, equals the input current flowing into the transistor. It is determined by the time constant C_{GS}/g_m.
g_m	Transconductance of a MOSFET
GBW	Gain Bandwidth product
I_D	Current flowing through the channel of a transistor. Defined as flowing from drain to source
K_p	$\mu_0 C_{ox}$
n	Emission coefficient
s	Laplace variable
V_E	Early voltage of a MOSFET
V_T	Threshold voltage of a transistor
V_{DSAT}	Saturation voltage of a transistor
V_{DS}	Drain source voltage, voltage difference between the drain and the source of a transistor
V_{GS}	Gate source voltage, voltage difference between the gate and the source of a transistor
v_{sat}	Saturation speed of the charge carriers in the channel of a MOSFET due to velocity saturation
AC	Alternating Current

BLA Best Linear Approximation

DC Direct Current

eSYSID electronic SYStem IDentification test-bench

LHP Left Half Plane

MIMO Multiple Input Multiple Output

ODE Ordinary Differential Equation

OTA Operational transconductance Amplifier

PLL Phase Locked Loop

RHP Right Half Plane

SISO Single Input Single Output

SNDR Signal to Noise and Distortion Ratio

SNR Signal to Noise Ratio

THD Total Harmonic Distortion

VCO Voltage Controlled Oscillator

Samenvatting

In [1] wordt de electronic SYStem IDentification test-bench (eSYSID) besproken, een programmeerbare test-bench die gebruikt kan worden in het experimenteel bevestigen van systeemidentificatie technieken, alsook in het aanleren van diezelfde technieken aan studenten.

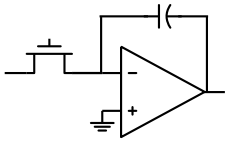
Het hart van de eSYSID test-bench is een digitaal programmeerbaar filter met prestaties die dezelfde zijn van chip tot chip. In deze thesis wordt het het ontwerp van de eSYSID grondiger bekeken en worden de eerste stappen in het ontwerp ervan gemaakt. Daarnaast wordt een analysemethode uitgewerkt die kan gebuikt worden in het verder ontwerp van de eSYSID.

In het eerste hoofdstuk wordt de hoogniveau architectuur van de eSYSID in detail bekeken. Er wordt gebruik gemaakt van de state-space voorstelling van (niet-)lineare systemen als basis voor de architectuur. Het basisprincipe bestaat erin om de matrixvermenigvuldigingen op een analoge manier uit te rekenen en zo een Multiple Input Multiple Output (MIMO) systeem met 3 ingangen en 2 uitgangen te implementeren. De mogelijke identificatie-experimenten evenals de calibratietechnieken die nodig zijn om de performantie robuuster en reproduceerbaarder te maken worden bekeken en er wordt een manier voorgesteld om de performantie ervan robuuster te maken tegen fouten in de parameters. Op het einde van dit hoofdstuk wordt een hoogniveau architectuur bekomen.

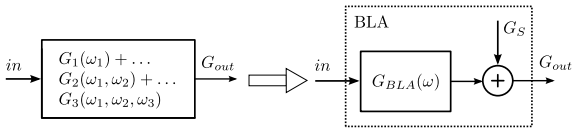
In het tweede hoofdstuk worden de mogelijke configuraties van programmeerbare filters in CMOS beschouwd. Elk van deze configuraties heeft zijn voor- en nadelen, deze worden onderzocht en de meest geschikte configuratie voor ons probleem wordt gekozen. MOSFET-C werd gekozen als techniek om de filter te realiseren. De basis bouwblokken in een MOSFET-C filter zijn een op-amp, capaciteitsbanken en een gekalibreerde transconductor (Zie Figuur 1). Om de performantie verder te verbeteren wordt voor een volledig differentiële uitvoering gekozen.

Het tweede deel van dit hoofdstuk bekijkt het ontwerp van de transconductor en het tuning netwerk dat nodig is om de transconductor te kalibreren. Op deze manier wordt een programmeerbaar filter bekomen dat robuust is tegen proces- en temperatuurvariaties.

Het derde hoofdstuk behandelt het ontwerp van de op-amp. De proces parameters worden uit het ingewikkeldere BSIM3V3 gehaald zodat met snelle handberekeningen een ontwerp bekomen wordt. Dat ontwerp wordt dan geverifieerd met simulaties en waar nodig aangepast. Er wordt een ontwerpstrategie gekozen die, startend van de specificaties tot een ontwerp komt. We beginnen met het toepassen van deze strategie op een simpel voorbeeld. Daarna beginnen we aan het echte werk: de volledig differentiële op-amp die gebruikt kan worden in de filter. We geven meer DC versterking aan de op-amp, kiezen een beter compensatie schema en onderzoeken de common-mode feedback schakelingen die mogelijk



Figuur 1: MOSFET-C integrator



Figuur 2: Toepassen van de BLA op een niet-lineair systeem

zijn. Het hoofdstuk eindigt met het eigenlijke ontwerp van de op-amp en de simulatieresultaten ervan.

Om de performantie van de op-amp in te schatten, en mogelijk te verbeteren, werd een niet-lineaire analysemethode uitgewerkt. Deze methode is het onderwerp van het vierde hoofdstuk. We maken gebruik van de Best Linear Approximation (BLA) om het niet-lineair gedrag van de op-amp te beschrijven. De BLA laat toe een niet-lineair systeem te benaderen door een lineair systeem en distorsie wordt beschouwd als ruis (Figuur 2). Door gebruik te maken van een random odd random fase multisinus kunnen we de distorsie gegenereerd door even en oneven niet-lineariteiten apart bepalen.

Vervolgens passen we technieken uit de klassieke ruisanalyse toe op die distorsie, namelijk het refereren naar de uitgang. Op die manier kunnen we de invloed van elke trap op de uitgang vergelijken en bekomen we een manier om het subsysteem dat verantwoordelijk is voor het genereren van het meest distorsie te bepalen. Het grote voordeel van deze analyse is dat ze op een op-amp in een feedback configuratie kan toegepast worden, zodat die feedback configuratie meegenomen wordt in de analyse.

Eerst wordt de analyse toegepast op een enkelzijdige tweetraps op-amp. De resultaten en bespreking ervan werden samengevat in een conference paper dat aanvaard werd voor publicatie. Het eerste deel van hoofdstuk is een kopie van dit paper.

Daarna gaan we iets dieper in op de fouten die gemaakt worden tijdens de analyse door het niet-lineaire instelpunt te gebruiken naast het gewone DC instelpunt. Dit leert ons veel over de beperkingen van de analyse. In het laatste deel van het hoofdstuk wordt de analyse uitgebreid en toegepast op een volledig differentiële op-amp met common-mode feedback. De uitbreiding is niet triviaal: de volledig differentiele op-amp is een MIMO systeem en wordt dus best met matrices beschreven. We passen de vergelijkingen voor het enkelzijdig geval aan zodat ze ook voor MIMO systemen werken. Tenslotte wordt de analyse toegepast op de volledige differentiële op-amp en de resultaten ervan worden getoond en besproken.

In het vijfde en laatste hoofdstuk worden de bekomen resultaten samengevat en wordt het mogelijk toekomstige werk opgesomd.

Chapter 1

High-Level Architecture

Experimental confirmation is the ultimate step in the development of a scientific method or theory. This is no different for system identification. The verification of the properties, advantages and disadvantages of an identification method requires a three step procedure

1. Design, realise and/or buy a system which is known as exactly as possible. This '*reference system*' aims at repeatability and well known, predictable operation, rather than top-notch performance.
2. Apply a collection of identification methods under test to the modelling of that '*reference system*'
3. Compare the parameters obtained with the identification method to the known parameters of the '*reference system*'

Up to now, the '*reference system*' usually boils down to a digital simulation. The limitations are clear: the prejudice of the experimenter is contained in the simulation and the nonidealities of the measurement set-up are not present in these simulations and can't be incorporated into the verification[1]. This is a problem, because measurements are the base of system identification. The downside is that including the measurements of a real system into the verification is hard.

Usually, custom-built systems act as the '*reference system*'. Because each research group uses their own custom-built systems, it is hard to compare the different methods developed by different research groups. A benchmark for identification methods is therefore a mandatory step towards the objective verification of identification methods under practically relevant operating conditions.

In this thesis, steps are taken in the development of such a standard test-bench. The goal is to have a design that can easily be distributed among the research groups. Thereto, the setup must be easily duplicated with a very highly repeatable performance. The advantages of such standardised test-bench are clear:

- Identification methods developed by different groups can be compared immediately;
- Experiments can easily be repeated and checked by other research groups with different measurement set-ups and conditions.

[1] For example, a method for time invariant systems will never encounter any drift. This can jeopardize the practical applicability of the method.

Because system identification is applied to a very wide range of systems, and since it wouldn't be very practical to build a test-bench for each type of system separately, the proposed test-bench needs some flexibility to represent a wide class of systems. A programmable analog signal processing module can implement a class of '*reference system*s'. Of course, the class is tuned to maximise the use of a class of systems in an experimental verification.

Specifications for a possible realisation of the test-bench are examined in [1] where a test-bench called eSYSID is proposed. The goal of this thesis is to contribute to the design and the development of the eSYSID.

In this chapter, we will look into the high-level architecture, determine whatever necessary systems and measurement set-ups can be represented by the architecture and try to obtain a general plan for the eventual chip. Considerations made here will lead to specifications for the blocks which are designed and analysed in the following chapters.

1.1 State-Space representation of linear and nonlinear systems

The eSYSID architecture proposes a test-bench which can be programmed to represent both linear and nonlinear systems by using linear dynamic and static nonlinear blocks. In this thesis, only the linear part will be considered. In order to build a system which can represent all the systems needed, we must first find a mathematical way of describing all of them. An Ordinary Differential Equation (ODE) should be a sufficiently accurate description. A general state-space representation of a system described by an ODE is given by

$$\begin{cases} \dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t), t) \\ \mathbf{y}(t) = g(\mathbf{x}(t), \mathbf{u}(t), t) \end{cases} \tag{1.1}$$

where $\mathbf{u}(t)$, $\mathbf{y}(t)$ and $\mathbf{x}(t)$ represent the inputs, outputs and the state vectors respectively. $\dot{\mathbf{x}}(t)$ represents the derivative of $\mathbf{x}(t)$ with respect to the time t. For linear dynamic systems, the functions $f(\cdot)$ and $g(\cdot)$ can be made using the following set of differential equations

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}(p,q)\mathbf{x}(t) + \mathbf{B}(p,q)\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}(p,q)\mathbf{x}(t) + \mathbf{D}(p,q)\mathbf{u}(t) \end{cases} \tag{1.2}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ represent the state-space matrices of the linear dynamics of the system. Since the eSYSID should be able to represent the complete class of linear dynamic systems, the coefficients of the state-space matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ should be made programmable. This programmability is obtained in two ways.

1. The parameters are variable and their value is made digitally programmable. This digital programmability is represented in the state-space equations by making all the parameters function of the discrete vector q. In a chip, programming precision is limited to a fixed amount of bits. Therefore q is discrete.

2. The state-space matrices are made tunable by a continuous value p in order to counter process and temperature variations of the chip. Because the limited amount of bits available in the q

vector, it lacks the precision needed to counter process variations. The p vector uses voltages instead of binary values to represent its value, so it can be considered as a continuous number.

Consider a Multiple Input Multiple Output (MIMO) linear system of 4^{th} order with 3 inputs and 2 outputs. This means **A** is a 4×4 matrix, **B** is a 4×3 **C** is a 4×3 matrix and **D** is a 4×3 matrix. With it, we can describe

4^{th} **order MIMO Linear Time Invariant (LTI) system** The LTI system is the most simple configuration. All inputs and outputs are used as signal lines. p is set to a constant value to counter process variations or, if the temperature changes, it is used to counter this change. q is used to set the values of the matrices to a constant value.

4^{th} **order MIMO Linear Parameter Varying (LPV) system** LPV systems can be described by varying the parameters p and/or q during the experiment. The configuration is the same as for the MIMO LTI system, but now the parameters p and q can be varied. For small precise variations, the parameter p can be adjusted around the value obtained to counter the process variations. For large but slow variations, q can be adjusted.

Errors In Variables (EIV) and Output Errors (OE) setup EIV and OE set-ups represent an important class of identification problems. Because noise at the measured input and output signals is considered, a real-world set-up is obtained. Filtered and/or correlated noise is added either to both inputs and outputs (EIV) or only to the output (OE). The set-up is shown in Figure 1.1. With our 4^{th} order system, a second order linear system with a second order noise model can be considered. Correlated filtered noise can be added by passing independent noise sources through a MIMO linear dynamic system. Therefore two of the inputs are used as noise inputs and one of the inputs as signal input.

This thesis considers a 4^{th} order programmable system. Before we look into the actual circuitry needed to implement this system, we will look into the sensitivity of the state space representation and try to improve its sensitivity by playing around with the parameters.

1.2 Robust state-space representation

In [2] it is argued that the eigenvalues of a matrix **A** are very sensitive to small changes to the non-zero values of that matrix. Tridiagonal matrices however, are relatively insensitive to changes

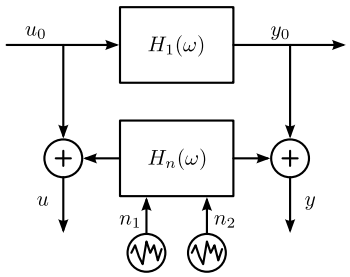


Figure 1.1: EIV set-up (Source: [1])

in the non-zero entries of the Matrix. In a state-space representation of a system, the eigenvalues of the system matrix $\mathbf{A}$ represent the poles of the described system [3]. Hence, the high sensitivity on the eigenvalues translates into a high sensitivity on the poles of the described system. Because the system has to be highly reliable and repeatable, high sensitivity to the element values cannot be tolerated. For a tridiagonal matrix however, this sensitivity is greatly reduced. Therefore we will apply a transformation which makes $\mathbf{A}$ tridiagonal.

It is possible to apply a nonsingular transform $\mathbf{T}$ on the state vector $\boldsymbol{x}$

$$\mathbf{T}\boldsymbol{z} = \boldsymbol{x} \tag{1.3}$$

This is called a similitude transformation. It results in the following state-space equations

$$\begin{cases} \mathbf{T}\dot{\boldsymbol{z}} = \mathbf{A}\mathbf{T}\boldsymbol{z} + \mathbf{B}\boldsymbol{u} \\ \boldsymbol{y} = \mathbf{C}\mathbf{T}\boldsymbol{z} + \mathbf{D}\boldsymbol{u} \end{cases} \Leftrightarrow \begin{cases} \dot{\boldsymbol{z}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} + \mathbf{T}^{-1}\mathbf{B}\boldsymbol{u} \\ \boldsymbol{y} = \mathbf{C}\mathbf{T}\boldsymbol{z} + \mathbf{D}\boldsymbol{u} \end{cases} \tag{1.4}$$

It is shown that we can find a state-space representation of the same system by transforming the state-space matrices as:

$$\begin{aligned} \mathbf{A}' &= \mathbf{T}^{-1}\mathbf{A}\mathbf{T} \\ \mathbf{B}' &= \mathbf{T}^{-1}\mathbf{B} \\ \mathbf{C}' &= \mathbf{C}\mathbf{T} \end{aligned}$$

In [2] a genetic algorithm is proposed to compute the tridiagonal system matrix starting from known poles. The Lanczos algorithm described in [4] can also be used to generate a tridiagonal matrix from a set of prescribed eigenvalues (poles). The parameters of $\mathbf{B}$ and $\mathbf{C}$ can be found by using the method described in [2]. If we apply the tridiagonalisation on the 4^{th} order system we want to build with 3 inputs and 2 outputs, we obtain the following set

$$\begin{aligned} \frac{d}{dt}\left(\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix}\right) &= \begin{bmatrix} a'_{11} & a'_{12} & 0 & 0 \\ a'_{21} & a'_{22} & a'_{23} & 0 \\ 0 & a'_{32} & a'_{33} & a'_{34} \\ 0 & 0 & a'_{43} & a'_{44} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} b'_{11} & b'_{12} & b'_{13} \\ b'_{21} & b'_{22} & b'_{23} \\ b'_{31} & b'_{32} & b'_{33} \\ b'_{41} & b'_{42} & b'_{43} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \\ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} c'_{11} & c'_{12} & c'_{13} & c'_{14} \\ c'_{21} & c'_{22} & c'_{23} & c'_{24} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \end{aligned} \tag{1.5}$$

with

$$\begin{aligned}
\boldsymbol{z} &= \mathbf{T}^{-1}\boldsymbol{x} \\
\mathbf{A}' &= \mathbf{T}^{-1}\mathbf{A}\mathbf{T} \\
\mathbf{B}' &= \mathbf{T}^{-1}\mathbf{B} \\
\mathbf{C}' &= \mathbf{C}\mathbf{T}
\end{aligned}$$

Implementing this matrix equation results in the system shown in Figure 1.2. This schematic clearly shows us the blocks we need to build to realise the filter. All parameters a_{ij}, b_{ij}, c_{ij} and d_{ij} must be made digitally programmable by the discrete parameter q and tune-able by the continuous parameter p. The system consists of 4 summer-integrators with either 5 or 6 signal inputs and 2 summer-amplifiers each with 6 inputs. In the following chapter, we look into the realisation of these blocks.

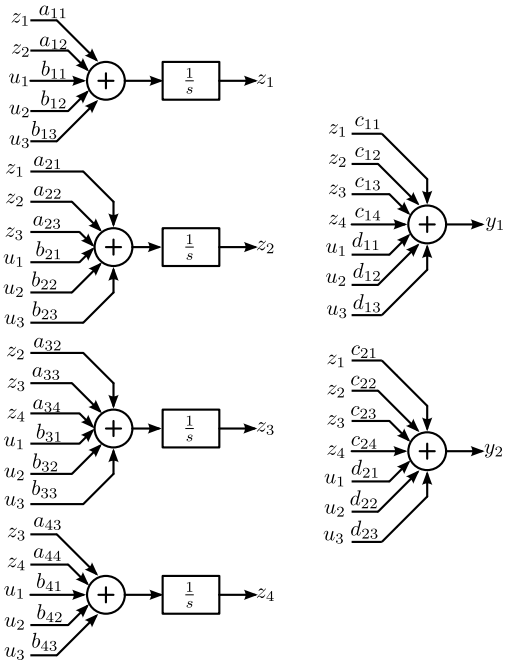


Figure 1.2: Total system to be built

Chapter 2

Filter Architecture

The high-level architecture for our chip was determined in the previous chapter. Now we will look into the possible circuits which exist in literature to realise the prescribed system. We will start by looking into the integrator, because it is the most difficult block to build and because the techniques used to build in the integrator can be used to build the summer-amplifier, which is the second block required on the chip.

Integrators have a $1/s$ frequency characteristic. This can be realised with an op-amp in a configuration as shown in Figure 2.1a. If the op-amp is considered ideal[1], we can assume node 1 is connected to virtual ground and that the op-amp draws no current. This allows us to assume the current determined by the voltage drop over the resistor ($i = V_{in}/R$) has to match the current flowing through the capacitor ($i = sC \cdot V_{out}$). Using these equations, we obtain the transfer function of the ideal integrator

$$\frac{V_{out}}{V_{in}} = -\frac{1}{sRC} \tag{2.1}$$

The frequency response of this transfer function is shown in Figure 2.1b.

In the first part of this chapter, we will consider the nonidealities which are introduced when we try to build this ideal integrator on a chip. It will result in a number of possible realisations for the integrator, each with its own advantages and disadvantages. From these realisations, the MOSFET-C integrator is chosen.

In the second part of the chapter, the MOSFET-C integrator is examined in more detail. This type of integrator consists of an op-amp, capacitors and a MOSFET in its linear region called a transconductor (Figure 2.2). Since the next chapter explains the design of the op-amp, the second part of the chapter will mainly deal with the design of the transconductor.

The gate voltage of the transistor in the transconductor will take the role of the parameter p from the previous chapter. Its main use is to compensate process and temperature variations on the chip and is generated with a tuning scheme which performs a calibration of the complete chip. Possible tuning schemes are discussed in the last part of this chapter.

[1] infinite input impedance, zero output impedance and an infinite gain over an infinite bandwidth

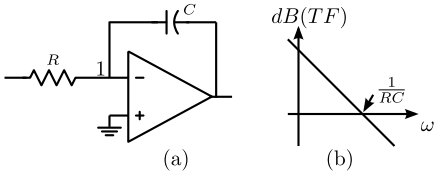


Figure 2.1: Integrator

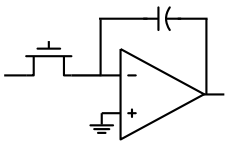


Figure 2.2: MOSFET-C Integrator

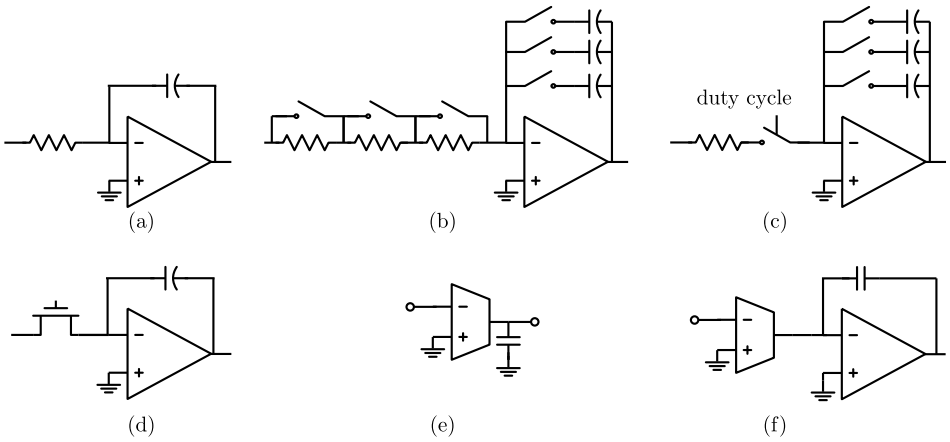


Figure 2.3: Different types of integrator. (a) The RC integrator (b) programmable/tunable RC integrator (c) switched-R integrator (d) MOSFET-C integrator (e) g_m-C integrator (f) OTA-C integrator.

2.1 Various integrator types

Looking at the previous chapter, we can derive some guidelines to choose the integrator realisation

- The integrator should be programmable with a discrete vector q and a continuous voltage p.
- Precision of about 12 to 14 bit should be obtained, so we will need a Signal to Noise and Distortion Ratio (SNDR) of 72 to 84dB .
- The needed bandwidth is limited up to a few kilohertz
- The performance of the integrator should be highly repeatable from chip to chip. A precise way of calibrating the integrator is necessary.

We shall now look how we can build an on-chip integrator which satisfies these specifications.

RC integrators The first way to construct an integrator is to use the ideal integrator from the introduction (Figure 2.3a), but now it's built with non-ideal components. Passive resistors can be made on-chip with polysilicon wires, capacitors in our technology of choice can be made with between two of the metal layers. Both these passive components are very linear. Due to this, the only source of distortion in the RC integrator is the op-amp itself which is placed in a feedback configuration. Due to the high gain of the op-amp in the feedback loop, high SNDR of up to 100dB can be obtained

[5] p. 568. The op-amp also determines the bandwidth limitation: the gain of the op-amp should be very high to suppress the distortion, so the bandwidth will be limited.

The biggest disadvantage for this integrator realisation is that on-chip absolute values of R and C are not well controlled. Therefore every resistor has to be tuned. This can be done using banks of capacitors and resistors. Figure 2.3b shows a possible realisation. The tuning range is limited though. When the parasitics of the switches are considered one needs to make a trade-off between tune-ability and parasitics. Tuning ranges up to 2% or 3% can be reached [6], which is insufficient for our purpose. A possible solution to this problem is to use a switched-R filter [7], shown in Figure 2.3c. By duty-cycling the resistor, it is possible to tune the value of the resistors in an analog way. Therefore a much larger tuning precision can be reached without losing linearity. A Total Harmonic Distortion (THD) of -90dB is reported [7][2]. The drawback of this technique is that this integrator is not a continuous-time integrator any more, so the problems of discrete-time systems like clock injection and noise folding plague this technique, resulting in lowered Signal to Noise Ratio (SNR) of 64dB for the complete filter.

MOSFET-C integrator A MOSFET in its linear region behaves like a resistor. Its resistance can be controlled with the gate voltage applied to that transistor, so the MOSFET in its linear region can be considered as a voltage controlled resistor. The integrator which uses this voltage controlled resistor is shown in Figure 2.3d. It is called the MOSFET-C integrator. Because it uses a voltage controlled resistor, the MOSFET-C integrator combines the high tune-ability of the switched-R filter with the continuous-time properties of the normal RC integrator. The nonlinear distortion will be higher than in the case of the RC integrator because the MOSFET is not an inherently linear device. With the topology shown in Figure 2.3d, a SNDR of 80dB can be reached [5]. The bandwidth will still be limited by the op-amp as for the RC integrator.

g_m-C filter In all previous cases, the bandwidth of the integrator is limited by the op-amp. To get rid of the bandwidth limitations of Active-RC and MOSFET-C integrators, the op-amp can be replaced by a high bandwidth Operational Transconductance Amplifier (OTA). The OTA takes an input voltage and converts it into an output current. This current is then integrated by a capacitor. The resulting integrator is shown in Figure 2.3e. The OTA is normally constructed with a differential pair to obtain a high bandwidth. To compensate for the low linearity of the differential pair, linearisation techniques are applied to the differential pair ([5] pp. 578-593). This type of continuous time filter is ideal for high frequency applications, but at low frequencies, their performance is lower compared to the other integrator types which contain a full op-amp: they don't have the high gain feedback mechanism to suppress distortion and noise. Because of this, these filters rarely achieve more than 60dB of SNDR [5].

OTA-C integrator We learned from the g_m-C filter that a transconductance g_m can be made with a linearised OTA. This transconductance can be used in an integrator with an op-amp instead of the resistance or linear-region MOSFET. The resulting topology is called an OTA-C integrator (shown in

[2]"measured total harmonic distortion of the filter at 2kHz for a switched-R filter built in a 0.18μm process with a 0.8V supply voltage. The input sinusoid was held at 0.6V Vpp differential".

Figure 2.3f). The limitations are similar to the ones found for the MOSFET-C integrator. The main difference is the fact that the transconductance g_m is made in an active way instead of a passive way.

Since the chip will be used in a frequency range from a few Hertz up to a few kilohertz, we don't need the big bandwidth offered by the g_m-C and since it's SNDR is very limited we won't choose for a g_m-C integrator. The performance of the chip should be accurate up to a very high level, so the reported 2% to 3% tune-ability of the active RC filter is too low. Although its SNDR is the best of all considered topologies, we won't choose for this integrator topology either. The MOSFET-C and OTA-C integrators remain. Their performance is similar, but we will choose for the MOSFET-C integrator, because its resistors are made in a passive way and with some modification to the MOSFETS, very high linearity can be obtained, something that is impossible to obtain with an OTA.

2.2 Differential versus single-ended systems

To improve the performance of the system, we shall use a fully differential topology. This increases the design complexity of the op-amp and the current consumption, but the advantages are huge:

Cancellation of even order nonlinearities If the inverse of a signal is presented to an even nonlinearity, its sign does not change. If each side of the differential system is presented to the same nonlinearity, the even nonlinearities cancel out when looking at the differential-mode. For a MOSFET-C integrator, where the nonlinear distortion is mainly even-order due to the quadratic behaviour of the linear-region MOSFETS, this cancellation of even-order nonlinearities is a huge benefit.

Greater immunity to disturbances Capacitive coupled disturbances can be made common-mode when the signal lines are put closely together. By looking at the differential-mode only, the disturbances disappear.

Greater power supply rejection ratio Disturbances to the power supply present themselves to the system as a common-mode disturbance. By looking only at the differential-mode, the influence of the power supply disturbance can be neglected. This property, combined with the disturbance immunity, is one of the main reasons why, these days, almost all analog circuitry on a mixed analog-digital chip is made differential.

Larger signal swing The maximum signal swing is mainly determined by the supply voltage. In a single-ended system, the signal swing can be one time the supply voltage. In a differential system, the maximal signal swing is twice the supply voltage.

Higher speed possible All op-amps have a differential pair as an input. In a single-ended circuit, the output of this differential pair has to be made single-ended. This is often done via a current mirror. This current mirror contains 2 gate capacitances and is therefore a slow circuit. It also introduces a zero at a frequency which is half the frequency of this pole. In a fully differential circuit, the mirror is not necessary, so the maximum speed that can be obtained with a differential amplifier is larger.

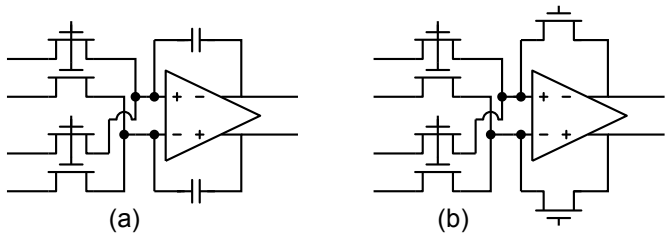


Figure 2.4: MOSFET-C summer integrator (a) and the MOSFET-C summer-amplifier (b)

The first three of these benefits are limited by the amount of mismatch between the elements in each path of the differential system. From now on, all circuits will be drawn in their differential topology. Creating this differential topology is done by just mirroring the single-ended system.

Figure 2.4a shows a fully differential MOSFET-C summer-integrator. The fully differential summer-amplifier made with linear-region MOSFETS is also shown (Figure 2.4b). These two building blocks can be used to build the complete architecture proposed in the previous chapter.

We will now study the operation and nonidealities of the MOSFET-C integrator in more detail. A fully differential MOSFET-C integrator consists of 3 parts: the fully differential op-amp, the capacitors in feedback and the transconductor which acts as resistors for the integrator. In the technology of our choice, the capacitors can be made easily on-chip between two metal layers. The op-amp and the transconductor are more challenging to design. The complete next chapter is dedicated to the design of the op-amp. The following sections of this chapter focus on the transconductor.

We start by summarising the basic definitions for the MOSFET transistor we will use throughout this thesis. Afterwards we will look into the transconductor.

2.3 The MOSFET

The Metal Oxide Semiconductor Field Effect Transistor (MOSFET) is the basic building block of the circuits in this thesis. We consider the MOSFET as a 4 terminal device. There are two variants: an NMOS and a PMOS variant, which have an n-doped and p-doped channel respectively. The symbolic representation of an NMOS is shown in Figure 2.5a. The four terminals are shown. They are called Gate, Source, Drain and Bulk with their respective voltage levels V_G, V_S, V_D and V_B as shown in Figure 2.5.

In its normal large signal mode of operation, the MOSFET has a conductive channel between source and drain which can be influenced by the gate voltage. Therefore, the most important voltages to describe the operating point of a MOSFET are the gate-source voltage V_{GS}, drain source voltage V_{DS} and the current flowing from drain to source I_D.

In this thesis, the model used to describe the operation of the MOSFET is called the quadratic model. It is a very simple model which is certainly not accurate enough for contemporary processes. However, the process used throughout this thesis is an 'old' 0.18μm CMOS process and the minimum gate length will rarely be used in the design. The long-channel approximation will therefore be satisfied in most cases. The quadratic model will yield results which are good enough to find initial values

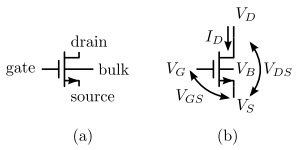


Figure 2.5: Definition of the MOSFET voltages and currents

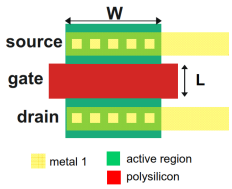


Figure 2.6: Lay-out of a MOSFET consisting of a single finger. (Source:[8])

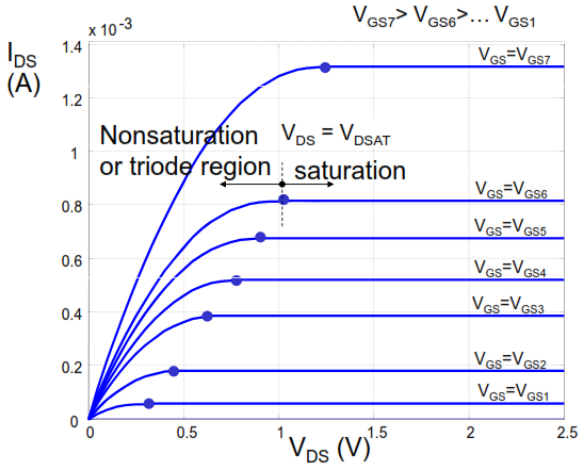


Figure 2.7: Equations described by the quadratic model. (Source:[8])

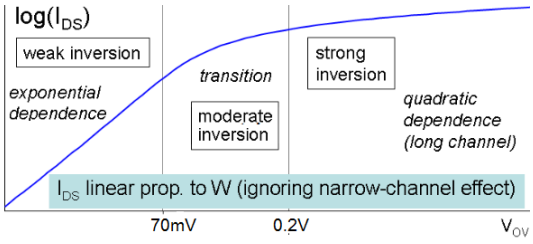


Figure 2.8: Drain Current I_D in function of the overdrive voltage. The three inversion levels can be clearly distinguished (Source:[8])

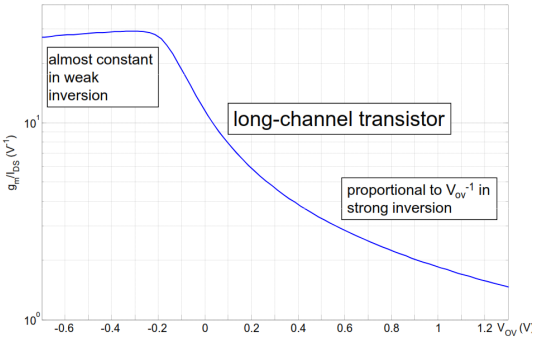


Figure 2.9: The efficiency g_m/I_D versus the overdrive voltage V_{ov} for a certain transistor (Source:[8])

through hand calculations. These initial values are then used in simulations which use a more complex BSIM3V3 model. The obtained initial values are then fine-tuned to take effects into account which are not included in the quadratic model.

The quadratic model consists of two equations, each for a certain region of operation called the linear region and saturation[3]:

Linear region ($\lvert V_{DS}\rvert < \lvert V_{DSAT}\rvert$)	Saturation ($\lvert V_{DS}\rvert > \lvert V_{DSAT}\rvert$)
$I_D = (W/L)\,\mu C_{ox}\left((V_{GS}-V_T)\,V_{DS} - \frac{V_{DS}^2}{2}\right)$	$I_D = \frac{1}{2}\,(W/L)\,\mu C_{ox}\,(V_{GS}-V_T)^2$

Plots of these functions for both regions are shown in Figure 2.7. The process parameters μ and C_{ox} will be discussed in the beginning of the next chapter. For now they can be considered as constants depending on the process.

The gate length L and gate width W are shown in Figure 2.6 for a simple lay-out. In the formulas they will always appear as the fraction W/L. Hence, the parameter W/L is one of the main design variables for MOSFET-based circuits. Two more operating parameters are introduced: V_T which is the threshold voltage and V_{DSAT} which is the saturation voltage.

Threshold voltage V_T The gate-source voltage V_{GS} required to produce an inversion layer. [9]

Saturation voltage V_{DSAT} The Drain-source voltage V_{DS} required for a transistor to be in the saturation region.

All expressions encountered so far are called the Large Signal expressions. It is custom in analog design to use the small-signal assumption and linearising the expressions around an operating point. This allows the designer to use the powerful linear time invariant framework in the design. One of the most important parameters that pops up during the linearisation is the transconductance g_m. It is defined as the change of the small signal current I_D to a change in V_{GS}.

$$g_m = \frac{\partial I_D}{\partial V_{GS}} \tag{2.2}$$

2.4 The MOSFET in saturation

We first consider the workings of the MOSFET in its saturation region. Remember that this means that $V_{DS} > V_{DSAT}$ and that, according to the quadratic model, the drain current is described by

$$I_D = \frac{1}{2}\,(W/L)\,\mu C_{ox}\,(V_{GS}-V_T)^2 \tag{2.3}$$

The expression $(V_{GS}-V_T)$ will, from now on, be called the overdrive voltage V_{ov}. Besides the W/L, The overdrive voltage is a second very important design parameter in CMOS design. It sets the inversion region in which the transistor operates. There are three operating regions: called the strong, moderate and weak inversion levels:

Strong Inversion $V_{ov} > 0.2$V: Due to the fact that the voltage applied to the gate is much larger than the threshold voltage, a big inversion layer will have formed under the gate. This makes that

[3]The absolute value of the voltages is taken to generalise the expressions for both NMOS and PMOS transistors

drift is the main mechanism behind the electron movement. The expression for the quadratic model is obtained under these assumptions.

Weak Inversion $V_{ov} < 70\text{mV}$: ([5] p. 19) In weak inversion, the main electron movement mechanism is diffusion. Due to this, the drain current I_D depends exponentially on the gate voltage. The quadratic model is certainly not valid in weak inversion.

Moderate Inversion $70\text{mV} < V_{ov} < 0.2\text{V}$: In this region, where the exponential behaviour of the weak inversion gradually turns into the quadratic behaviour of strong inversion.

The drain current I_D in function of the overdrive voltage for a given transistor is shown on a logarithmic scale in Figure 2.8. The exponential relation in weak inversion on a logarithmic scale corresponds to the linear dependence seen in the figure. The quadratic dependence of strong inversion and the transition in moderate inversion are also clearly shown in the figure.

The saturation voltage V_{DSAT} lowers when the overdrive voltage decreases. This means that the voltage range where the transistors are in saturation increases. Moderate and weak inversion therefore become more and more important for designs made in processes with a lower supply voltage. As we use an old technology, we are not plagued by a very limited supply voltage[4]. Therefore we are not obliged to use the weak and moderate inversion levels.

There is a good reason to use the weak and moderate inversion levels, even in the old technology of our choise. We can define the efficiency of a transistor by the ratio between the transconductance g_m and the current I_D necessary to obtain that g_m. This ratio ${}^{g_m}/_{I_D}$ versus the overdrive voltage is shown in Figure 2.9. A transistor in weak inversion is the most efficient. The problem with transistors biased in weak inversion is that they need to be very big (a gate width W of a few meters if one is not careful). Therefore, we will bias most of the transistors in this thesis on the edge of the strong inversion region where $V_{ov} = 0.2\text{V}$, where the quadratic model still works, but where the transistor is quite efficient. Only if the need arises, the transistors will be biased differently.

2.5 The MOSFET in its linear region

The transconductor consists of MOSFETS in their linear region. A simple concept in itself, but since its operation is critical to the performance of the whole integrator, it will be thoroughly examined. Using the quadratic model from the previous section and assuming the device is in the linear region ($V_{DS} < V_{DSAT}$), the drain current I_D is given by

$$I_D = ({}^{W}/_{L})\,\mu C_{ox}\left((V_{GS} - V_T)\,V_{DS} - \frac{V_{DS}^2}{2}\right) \tag{2.4}$$

Applying this formula to the situation shown in Figure 2.10, and looking at the linear term only, we obtain the following expression

$$I_D = ({}^{W}/_{L})\,\mu C_{ox}\,(V_G - V_Q - V_T)\,(V_D - V_S) \tag{2.5}$$

[4]Due to the fact a fully differential system is considered, the maximal voltage swing is even twice the supply voltage!

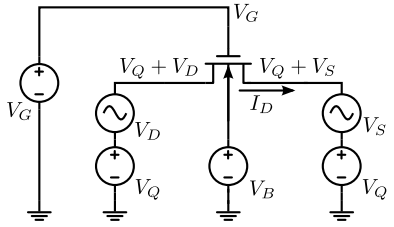


Figure 2.10: Definition of the voltages and currents around the MOSFET.

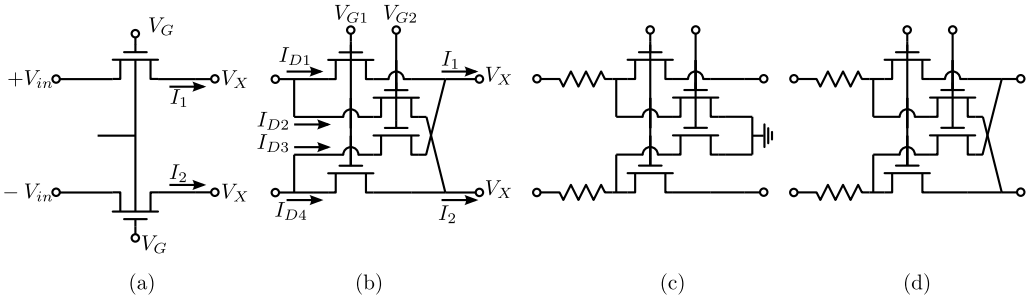


Figure 2.11: Different types of transconductors. (a) two-transistor transconductor, (b) four-transistor transconductor and (c-d) R-MOSFET transconductor.

with V_Q the signal ground, a set DC voltage around which the differential signal is applied. Throughout this thesis the signal ground will be half the supply voltage, so $V_Q = 0.9V$. The voltage for the signal ground can be chose differently for the transconductor in order to increase its tuning range[10]. Using this expression allows us to describe the working of the transconductor. We will start with the normal differential realisation: the two-transistor transconductor.

2.5.1 Two-transistor transconductor

If we apply equation (2.5) from the previous paragraph to the situation described in Figure 2.11a, where the two transistors are matched, we obtain an expression for the current in both branches of the differential system. Because the transconductor is attached to an op-amp, we can assume both outputs are at the same voltage (virtual ground). We call this voltage V_X. This yields

$$\begin{cases} I_1 = (W/L)\,\mu C_{ox}\,(V_G - V_Q - V_T)\,(V_X - V_{in}) \\ I_2 = (W/L)\,\mu C_{ox}\,(V_G - V_Q - V_T)\,(V_X - (-V_{in})) \end{cases} \tag{2.6}$$

If we consider the differential current $i_1 - i_2$ we find

$$I_1 - I_2 = (W/L)\,\mu C_{ox}\,(V_G - V_Q - V_T)\,(2V_{in}) \tag{2.7}$$

We can define the differential resistance of the transconductor as the ratio between the differential voltage $V_{diff} = 2 * V_{in}$ and the differential current $I_{diff} = I_1 - I_2$.

$$\frac{V_{diff}}{I_{diff}} = R_{diff} = \frac{1}{(W/L)\,\mu C_{ox}\,(V_G - V_Q - V_T)} \tag{2.8}$$

In this structure, the even order nonlinearity cancels due to the fact a differential system is considered. The odd nonlinearity remains, but is considered to be sufficiently small for most applications [11].

2.5.2 Four-transistor transconductor

In [12] a transconductor is proposed with four matched transistors instead of two. It is called the Czarnul-Song transconductor and it is shown in Figure 2.11b. Applying the Kirchoff current law to the output nodes, we obtain

$$\begin{cases} I_1 = I_{D1} + I_{D3} \\ I_2 = I_{D2} + I_{D4} \end{cases} \tag{2.9}$$

The four-transistor transconductor is also attached to an op-amp. Therefore its two output nodes are again put at a voltage V_X. There are two control voltages now, called V_{G1} and V_{G2}. Applying equation (2.5) to the current situation, yields

$$\begin{cases} I_1 = (W/L)\,\mu C_{ox}\left[(V_{G1} - V_Q - V_T)\,(V_X - V_{in}) + (V_{G2} - V_Q - V_T)\,(V_X + V_{in})\right] \\ I_2 = (W/L)\,\mu C_{ox}\left[(V_{G2} - V_Q - V_T)\,(V_X - V_{in}) + (V_{G1} - V_Q - V_T)\,(V_X + V_{in})\right] \end{cases} \tag{2.10}$$

$$I_2 - I_1 = 2V_{in}\,(W/L)\,\mu C_{ox}\left[V_{G2} - V_{G1}\right] \tag{2.11}$$

$$R_{diff} = \frac{1}{(W/L)\,\mu C_{ox}\left[V_{G2} - V_{G1}\right]} \tag{2.12}$$

V_Q and V_T disappear from the expressions, making the resistance more tolerant to substrate noise [6][5]. The second advantage of this type of transconductor is that the resistance can be made negative by making V_{G1} larger than V_{G2}. This property is used in active mixers and we will use it in the R-MOSFET transconductor.

2.5.3 Resistors combined with a transconductor: R-MOSFET transconductor

Since the nonlinear contributions depend on the voltage across the transconductor, it is best to minimise this voltage. By placing a normal resistor in series with the transconductor, part of the voltage drop occurs there and, because the resistor is a perfectly linear device, no distortion is introduced. In [16], two R-MOSFET topologies are discussed. They are shown in Figure 2.11c and d. The first one (Figure 2.11c) can divert some of the current coming from the resistors towards ground. This way, the maximum resistance of the total structure can be increased. The second R-MOSFET transconductor (Figure 2.11d) uses the 4 transistor transconductor explained earlier. Because the 4 transistor transconductor allows to present a negative differential resistance, its use in the R-MOSFET transconductor allows to increase the tuning range of the whole structure.

[5] Many authors conclude complete nonlinearity cancellation happens in this four-transisor transconductor [13, 14, 15], but tests using harmonic balance simulations have shown no improvement of the nonlinear behaviour of the circuit.

Choosing one of the four proposed transconductors and designing them is not considered in this thesis. The design of the transconductor should be incorporated into the global (nonlinear) optimisation of the whole integrator using the nonlinear analysis method that will be described in Chapter 4. On the other hand, influence of mismatches and variance of process parameters must be incorporated into the design, something that is not considered here.

2.6 Automatic tuning

The process parameters of a chip depend on a lot of factors during the manufacturing of the chip. Since the link between the resistance presented by the transconductor and it's control voltage is determined by parameters such as μ and C_{ox} (which will vary from chip to chip and in function of temperature). The needed control voltage cannot be predicted in advance and applied to the transconductor with a normal voltage source.

In the system architecture, this need for a calibration signal was represented by the continuous vector p. The value of p has to be tuned to get the behaviour of the circuit identical to an exactly known reference and adapt p then so that the output of the comparison circuit matches the reference. The circuit that performs this comparison and which tunes p to compensate errors is called the tuning scheme. Several tuning schemes are suggested in [13]. They can be divided into two categories:

Offline tuning The filter is taken off-line and its characteristics are compared to a reference in order to be tuned. This method can achieve very good results when all variations to the filter are time-independent or slow compared to the time between two tuning actions. Because the actual filter is tuned, mismatches don't play a role.

Online tuning A copy of the elements is used to make a reference circuit on-chip which is tuned online. All the other elements in the filter are assumed to be identical, so they are all tuned this way. Mismatches between the copied element and the working elements come into play and place a limit on the precision of the tuning.

Due to the fact that the filter is completely programmable, an offline tuning scheme is difficult to implement because the circuit then depends on the programmed state of the filter. The fact that the system has to be taken off-line to perform the calibration also hampers the operation too much. Therefore, an online tuning will be used.

Even with online tuning, the whole filter can be copied and fed with a reference signal. Feedback can be applied to the copied filter as to predict the reference values needed for the working filter (Figure 2.12a). Again, problems arise due to the programmability of the filter. Therefore we will tune smaller pieces of the filter.

The standard transconductor can be copied and its resistance can be compared to a precise external resistor (Figure 2.12b). This tuning scheme has the advantage of being very simple, but it cannot take nonidealities of the op-amps into account.

The tuning scheme of choice for our system uses a Phase Locked Loop (PLL). A Voltage Controlled Oscillator or VCO can be made with two MOSFET-C integrators (Figure 2.13) and its oscillation frequency can be adjusted with the control voltage of the transconductors.

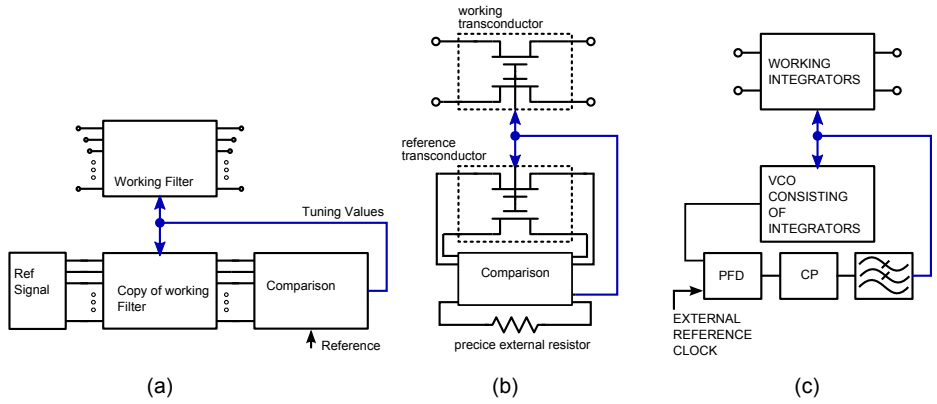


Figure 2.12: Automatic online tuning schemes. (a) a copy of the full filter is tuned (b) a reference transconductor is tuned (c) PLL tuning of the integrators

The phase of the output signal of this VCO is compared to an external reference signal generated with a crystal. A Phase Frequency Detector (PFD in Figure 2.12c) built with standard logic cells can perform this phase comparison (Figure 2.14 and [17]). The output of the PFD is sent through a charge pump (CP in Figure 2.12c) and is inserted into a loop filter. The output of the loop filter is then used to control the VCO and is used as the reference voltage for the filter.

The oscillation frequency of the VCO is tuned with the accuracy of a crystal oscillator. Hence, the time constant of both the VCO and all other integrators are set with a high accuracy.

Since the VCO is made with exact copies of the integrators used in the filter, this tuning scheme can compensate for the parasitic capacitances and nonidealities of the op-amp, so its performance will be a lot better than the one which uses only the transconductor as a reference element.

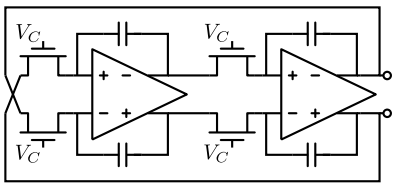


Figure 2.13: Voltage Controlled Oscillator (VCO)

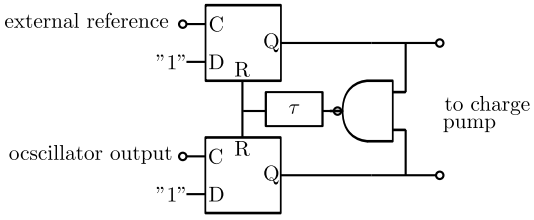


Figure 2.14: Phase Frequency Detector (PFD)

Chapter 3

Op-amp design

The op-amp is the basic building block in almost all analog devices and our programmable system is no different. Asides from some influence from the transconductors, the system performance will be mainly determined by the quality of the op-amp. Therefore we will spend a lot of time designing the op-amp.

The op-amp design starts on paper. Quick hand calculations and approximate expressions allow us to obtain an approximate sizing of the transistors and the compensation capacitances. The design found with the hand calculations is then simulated and tuned to account for the approximative nature of the hand calculations.

For the hand calculations, we first need to determine some process parameters. Second, the design strategy for our op-amp is explained and applied to a simple example. Afterwards, a fully differential op-amp architecture and a common-mode feedback circuit are chosen. Finally, the design plan is applied to this op-amp and the simulation results are shown and discussed.

3.1 One-page MOS model

In [5] p. 38, formulas are given to perform hand calculations for design, based on the quadratic model. These will be the formula's we use throughout the design.

$$I_D = \frac{K_p}{2n}\frac{W}{L}\left(V_{GS}-V_T\right)^2 \tag{3.1}$$

$$g_m = \frac{K_p}{n}\frac{W}{L}\left(V_{GS}-V_T\right) = \frac{2I_D}{V_{GS}-V_T} \tag{3.2}$$

$$r_o = \frac{V_E L}{I_{DS}} \tag{3.3}$$

$$f_T = \frac{1}{2\pi}\frac{3}{2n}\frac{\mu}{L^2}\left(V_{GS}-V_T\right) \quad or \approx \frac{v_{sat}}{2\pi L} \tag{3.4}$$

Abbreviation	Units	Name
K_p	$\mathrm{A/V^2}$	$\mu_0 C_{ox}$
μ_0	$\mathrm{m/Vs}$	mobility
C_{ox}	$\mathrm{F/m^2}$	gate capacitance
n		emission coefficient
v_{sat}	$\mathrm{m/s}$	saturation speed
V_E	$\mathrm{V/m}$	Early voltage

For the 0.18μm process, which will be used for the design, only a BSIM3V3 model of the transistors is available. In the BSIM3V3 model [18], only v_{sat} and μ_0 are specified directly. In the hand calculations,

we will use the "MIXEDMODE twin-well 1.8V transistor" model in the TT process corner.

Parameter	NMOS	PMOS	Unit
v_{sat}	$7.158 \cdot 10^4$	$5.34 \cdot 10^4$	$^{m}/_{s}$
μ_0	$3.141 \cdot 10^2$	$1.145 \cdot 10^2$	$^{cm^2}/_{Vs}$

Table 3.1: Parameters and their value as specified in BSIM

The other parameters have to be derived. We start with C_{ox}. The following formula can be used to determine it

$$C_{ox} = \varepsilon / T_{ox} \tag{3.5}$$

with T_{ox} the oxide thickness and ε the permittivity of the oxide. For SiO_2, the ε_r is 3.9. Since the thickness and permittivity are the same for NMOS and PMOS transistors, the C_{ox} parameter is equal. In our case T_{ox} equals 4.2nm, and hence $C_{ox} = 0.0082^{F}/_{m^2}$.

With C_{ox} and the given mobility, we calculate K_p for the PMOS and NMOS transistors

$$K_{p,NMOS} = \mu_0 C_{ox} = 2.58 \cdot 10^{-4\,A}/_{V^2} \qquad K_{p,PMOS} = 9.39 \cdot 10^{-5\,A}/_{V^2} \tag{3.6}$$

In order to calculate the current I_{DS} using the quadratic model, the effective mobility has to be used. Because the expressions for the effective mobility in the BSIM model are very complicated, we use the emission coefficient as the correction factor. For silicon, this emission factor n is equal to 1.3. This gives us the approximate parameters of the CMOS process we will use.

Abbreviation	Units	NMOS	PMOS
K_p	$^{\mu A}/_{V^2}$	258	93.9
n		1.3	1.3
v_{sat}	$^{m}/_{s}$	$7.158 \cdot 10^4$	$5.34 \cdot 10^4$
C_{ox}	$^{fF}/_{\mu m^2}$	8.2	8.2
μ_0	$^{cm^2}/_{Vs}$	$3.141 \cdot 10^2$	$1.145 \cdot 10^2$

Table 3.2: Summary of the process parameters

3.2 Design strategy for a Miller compensated op-amp

Now that we know the process parameters, we can perform hand calculations to get an approximate sizing of the transistors in function of the wanted g_m or I_{ds}. The next part of the chapter will deal with determining which g_m and/or I_d are necessary for each transistor in an op-amp. This step is called the design plan. The design plan starts with an op-amp architecture and specifications for the op-amp. From these, the required g_m and I_d of each transistor is obtained. Using the one-page MOS model from the previous section, we then obtain a sizing for each transistor.

In [5] a technique for the design of two-stage amplifiers is described. A generic Miller compensated 2-stage op-amp can be represented as in Figure 3.1. Each stage is modelled by a voltage controlled current source with a transconductance g_m. Then the most important capacitors for the design are

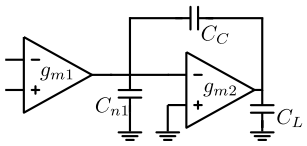


Figure 3.1: two-stage op-amp

added: the load capacitance C_L, the compensation capacitance C_C and the parasitic capacitance at the internal node C_{n1}. This capacitance is mainly determined by the input capacitance of the second stage. The gain-bandwidth product (GBW) of this circuit is given by[5]

$$GBW = \frac{g_{m1}}{2\pi C_C} \tag{3.7}$$

and the frequency of the non-dominant pole is given by.

$$f_{nd} = \frac{g_{m2}}{2\pi C_L} \frac{1}{1 + \frac{c_{n1}}{c_C}} \tag{3.8}$$

These two equations describe the op-amp. C_L is specified in the specifications for the op-amp and C_{n1} is fixed by the input capacitance of the second stage. This leaves us variables to play around with: g_{m1}, g_{m2} and C_c. We have two expressions: (3.7) and (3.8) and three variables, so we can choose the value of one of them.

Because the value of C_C is the most limited ($3 \cdot C_{n1} < C_C < C_L$) we pick C_C freely and let it determine the values of g_{m1} and g_{m2}. The ratio between C_L and C_C is called α from now on.

$$\alpha = \frac{C_L}{C_C} \tag{3.9}$$

If only the gain bandwidth GBW is specified, we have to choose the location of the second pole too. By prescribing the wanted phase margin, the location of the non-dominant pole f_{nd} to the GBW is fixed. For a phase margin of 63 degrees for example, the non-dominant pole must lie at a frequency which is the double of the GBW. We use the factor γ to denote the ratio between the GBW and the non-dominant pole frequency f_{nd}.

$$\gamma = \frac{f_{nd}}{GBW} \tag{3.10}$$

The last thing we need to solve from (3.8) is the value of the input capacitance of the second stage C_{n1}. Since we don't know this value yet, we make an educated guess. We use the parameter β to denote the ratio between C_c and C_{n1}.

$$\beta = \frac{C_C}{C_{n1}} \tag{3.11}$$

We shall assume a β of about 3. The parameters α, β and γ can now be used to design the two-stage op-amp. We will illustrate the design method on a simple Miller op-amp to show how it works.

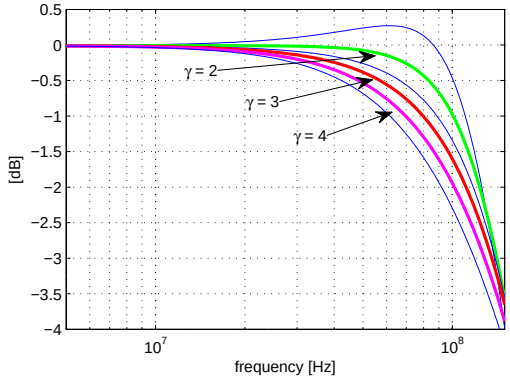


Figure 3.2: Frequency response of a two-pole system with one dominant pole and a second pole at γ times the GBW. The system is placed in unity feedback.

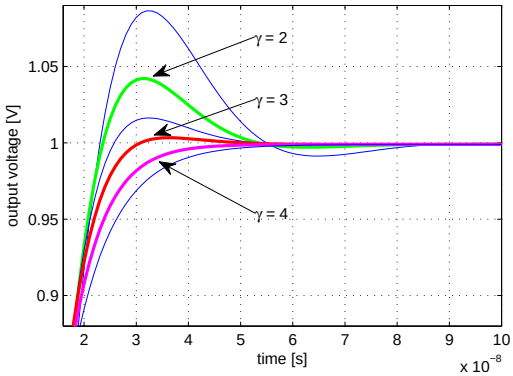


Figure 3.3: Step response of the same two-pole system in unity feedback for varying γ.

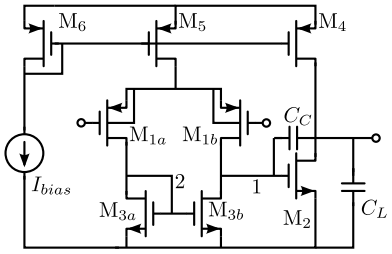


Figure 3.4: The Miller op-amp

3.3 Design example: Miller op-amp

The design method will be applied to the op-amp shown in Figure 3.4. This single-ended op-amp can be used in the loop filter and the charge pump of the PLL-based tuning network. The Miller op-amp we consider is a simple design, so it serves as an excellent example to show how the design plan works, before we apply it to more difficult op-amp architectures. We want a GBW of 200 MHz and we want to be able to drive a load capacitance of 5pF.

We begin by choosing the design parameters α, β and γ. Since this op-amp can be used as an all-purpose component throughout the chip, we don't know the exact feedback configuration the op-amp will be placed into. We will have to assume the worst: unity feedback. If we consider the op-amp as a two-pole system, we can impose it to be a critically damped system under unity feedback. This would mean the nondominant pole should lie at a frequency which is 4 times the GBW, or a $\gamma = 4$ [19]. Placing the nondominant pole that far from the GBW frequency requires a high g_m in the second stage, resulting in a large current consumption in the second stage. This is not something we want.

Figure 3.2 shows the amplitude characteristic of a two-pole op-amp in unity feedback with varying γ. There it is shown that γ can be lowered without introducing peaking in the frequency characteristic of the op-amp in unity feedback. $\gamma = 2$ is the minimum value for which no such peaking occurs. We shall choose $\gamma = 3$ so that the op-amp is still robust against process variations.

We choose $\alpha = 3$ and $\beta = 3$ following the arguments made in [5] pp. 188-194. Solving equations (3.7) and (3.8) gives us g_{m1} and g_{m2}

$$g_{m1} = 2\pi GBW \cdot C_L \frac{1}{\alpha} = 2,1\text{mS} \tag{3.12}$$

$$g_{m2} = 2\pi GBW \cdot C_L \left(1 + \frac{1}{\beta}\right) \gamma = 25,1\text{mS} \tag{3.13}$$

Since we want an efficient op-amp, we bias its transistors on the edge of strong inversion with an overdrive voltage ($V_{ov} = V_{GS} - V_T$) of 0.2V, which corresponds to a of ${}^{g_m}/_{I_d}$ of 10 . Since we know the current flowing through the transistors, it is possible to size them using our simple MOS model (3.1)

$$\left(\frac{W}{L}\right)_1 = {}^{I_{ds1}}\Big/_{\frac{K_{p,PMOS}}{2n} V_{ov}^2} = 145 \qquad \left(\frac{W}{L}\right)_2 = {}^{I_{ds2}}\Big/_{\frac{K_{p,NMOS}}{2n} V_{ov}^2} = 633 \tag{3.14}$$

The current through transistors 1 and 2 is now determined, it is now possible to choose the overdrive voltage of the remaining transistors and determine their ${}^{W}/_{L}$ using the same method as for transistors 1 and 2.

$$\left(\frac{W}{L}\right)_3 = {}^{I_{ds4}}\Big/_{\frac{K_{p,PMOS}}{2n} V_{ov}^2} = 53 \qquad \left(\frac{W}{L}\right)_4 = {}^{I_{ds1}}\Big/_{\frac{K_{p,NMOS}}{2n} V_{ov}^2} = 1739 \tag{3.15}$$

$$\left(\frac{W}{L}\right)_5 = {}^{2 \cdot I_{ds1}}\Big/_{\frac{K_{p,PMOS}}{2n} V_{ov}^2} = 290 \tag{3.16}$$

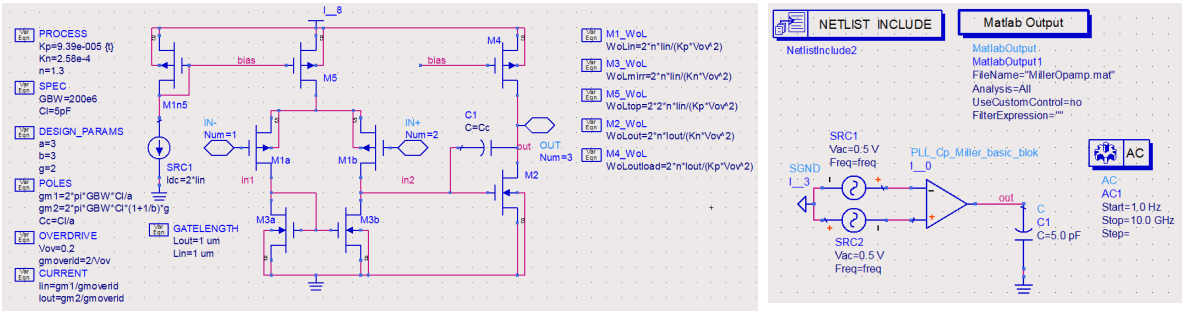
Only the length of the transistors remains to be determined. We don't take the minimum gate length of $0.18\mu m$ as the intrinsic gain of the transistors decreases with the gate-length and hence the gain of the amplifier. Including the length in the design is normally done by placing the parasitic poles at a sufficiently high frequency. Using approximate expressions for the remaining poles in function of the transition frequency f_T and the expression (3.4) from the one-page MOS model, the location of these poles can be set and a gate length can be obtained ([5] p. 183).

This method becomes difficult to use for more complicated circuits. Instead, we will use simulations to determine the optimal gate length for our op-amp. The simulator of our choice is Advanced Design System (ADS)[1] of Agilent.

The simulation set-up is shown in Figure 3.5. The op-amp is put into a cell which contains all the transistors. Normally the bias circuit is placed into another cell, because when many copies of the op-amp are present on one chip, they can share the bias circuit. To keep things simple, this is not done here. Around the op-amp cell, a test-bench is built, which contains the excitation sources and simulation components. This way, the design and testing of the op-amp are independent of each other and the op-amp can immediately be inserted in a larger design.

We use the power of ADS and engineering intuition to determine the gate length: manual tuning. By inserting the width of the transistors in the simulation as a function of the length and the ${}^{W}/_{L}$, it becomes possible to tune the length, without influencing the operating point of the transistor too much. Using the tuning function which is built into ADS allows then to quickly simulate the FRF of

[1] http://www.home.agilent.com



(a) Schematic of the Miller op-amp in ADS

(b) Simulation set-up

Figure 3.5: Miller op-amp simulation set-up

the op-amp and determining the influence of the parasitic poles on the op-amp.

Because transistors M_4, M_5 and M_6 share their biasing voltage, we will give them the same gate-length L_{load}. The same goes for transistors M_2, M_{3a} and M_{3b}. Their gate-length will be called L_{out}. The gate-length of M_1 (L_{in}) can be tuned independently. For convenience, L_{load} and L_{out} are chosen to be equal. This means only two parameters have to be tuned: L_{in} and $L_{out} = L_{load}$.

The tuning starts with a large gate length of 1μm for all the transistors. At this length, the drain bulk capacitance of the load transistor of the output stage M_4 is too large and it shifts the frequency of the nondominant pole to lower frequencies, killing the phase margin. Decreasing the parameter L_{out} to 0.42μm solves the problem. The gate-length of the input transistor is kept at 1μm.

With the gate lengths determined, we can perform some final simulations to determine the performance of our op-amp. The simulation set-up is placed in the test-bench for the op-amp and is shown in Figure 3.5b. The results of the simulation are exported to MATLAB[2] to generate nice figures. This is done with the '*MatlabOutput*' component in ADS. The result of the AC analysis is shown in Figure 3.6. The GBW of the op-amp is 138MHz with a phase margin of 61°. A DC gain of 60dB was obtained.

As a final test, we simulate the step response of the op-amp placed in unity feedback. This way, we can compare the obtained step response to the theoretical shown in Figure 3.3. The result of this transient simulation is shown in Figure 3.7. The overshoot corresponds to a γ between 2 and 3, the value we wanted.

Looking at the step response more closely (Figure 3.8), we notice two strange things at the point where the step occurs. First the output of the op-amp makes a small jump following the input. This can be explained by capacitive coupling between input and output. After the small jump, the output goes into the wrong direction! This effect is due to a zero in the right half of the complex plane and is certainly unwanted. The origin of this zero will be explained next.

3.4 Right half plane zero

A feed-forward path exists around the second stage of the op-amp due to the compensation capacitor. The result of this feed-forward path is a zero, placed in the right half of the complex plane (RHP zero). Looking at the open loop FRF of the op-amp, an RHP zero acts as a normal zero: it adds

[2]http://www.mathworks.nl/products/matlab/

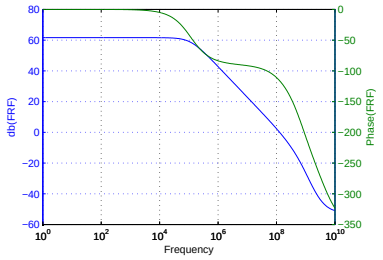


Figure 3.6: Results of the open-loop AC analysis in ADS

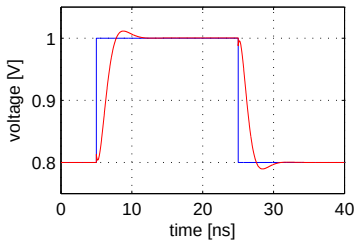


Figure 3.7: Results of the transient analysis in ADS

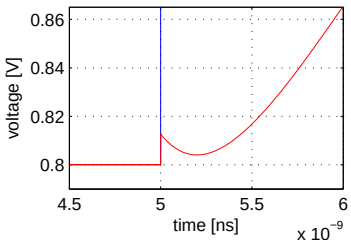


Figure 3.8: detail of the step response of the Miller op-amp

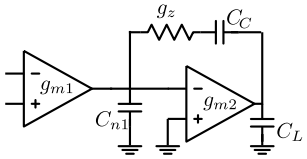


Figure 3.9: Miller op-amp with RHP zero compensation

20dB/decade to the slope of the amplitude. For the phase though, it acts as a pole: it adds a phase lag of 90 degrees. This kills the performance of an op-amp, because the GBW increases while the phase margin decreases. Its effect on the step response of the a closed loop system was shown before. The frequency of this RHP zero is[19]

$$f_{RHP} = \frac{g_{m,out}}{2\pi C_C} \tag{3.17}$$

It will lie beyond the GBW since $g_{m,out}$ is larger than $g_{m,in}$. A first way to compensate this RHP zero is by adding a resistor in the compensation feedback path (as shown in Figure 3.9). By adding a resistor with conductance g_z in the feedback path, the time constant associated with the RHP zero shifts to[19]

$$f_{RHP} = \left(2\pi C_C \left(\frac{1}{g_{m2}} - \frac{1}{g_z}\right)\right)^{-1} \tag{3.18}$$

By playing around with the value of g_z, we can place the zero where we want. We can either make it 'disappear' by choosing g_z equal to $g_{m,out}$. It is difficult however to match a resistor to a g_m because of process variations. Alternatively, we can place the zero in the left half plane (LHP) by choosing g_z smaller than $g_{m,out}$. We can cancel the nondominant pole by placing the zero at γ times the GBW[3]. Combining the two limits we found, we get an expression for the conductance needed in the compensation path

$$2g_{m1} < g_z < g_{m,out} \tag{3.19}$$

[3]Note that we cannot use this to place the nondominant pole at a frequency lower than the GBW and then compensate it by placing the LHP zero on top of it. if we do this, due to mismatches, a pole-zero doublet can occur. This introduces a slow settling mode in the step response, killing the settling time. [5] p. 84.

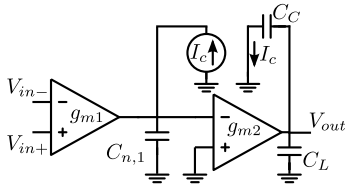


Figure 3.10: Ahuja compensated two-stage op-amp

3.5 Ahuja compensated op-amp

Another way of cancelling the influence of the RHP zero is by inserting a buffer in the feedback path. When a current buffer is placed in the compensation feedback path, the compensation technique is called Ahuja compensation. Asides from the RHP zero, the biggest disadvantage of a Miller compensated op-amp is the fact that the location of the non-dominant pole is determined by g_{m2}/C_L (see (3.8)). Placing this pole at a high frequency requires a lot of current in the second stage to raise the g_{m2}. If we look at the Miller compensated op-amp from the example, we see the second stage consumes about 8 times more current than the first one. Placing a current buffer in the feedback path allows more control of the location of the nondominant pole, so that we can lower the current consumption of the second stage. In [19] the location of the non-dominant pole of the Ahuja-compensated op-amp is determined

$$f_{nd} = \frac{g_{m2}}{2\pi(C_C + C_L)} \frac{C_C}{C_{n1}} \tag{3.20}$$

We can recycle the design strategy for the Miller compensated op-amp, but now use (3.20) for the location of the non-dominant pole. Choosing α, β and γ and solving both equations to g_m gives

$$g_{m1} = \frac{2\pi GBW \cdot C_L}{\alpha} \tag{3.21}$$

$$g_{m2} = 2\pi GBW \cdot C_L \frac{\gamma}{\beta}\left(\frac{1}{\alpha} + 1\right) \tag{3.22}$$

For a transistor biased in the strong inversion region, g_m depends linearly on I_d. We can therefore determine the ratio between the currents flowing through the first and second stage by dividing the g_m of the stages:

$$\frac{g_{m2}}{g_{m1}} = \frac{\gamma(1+\alpha)}{\beta} \approx 4 \tag{3.23}$$

This ratio is a lot smaller than the ratio of 12 for the Miller-compensated op-amp. We can thus decide that the Ahuja compensated op-amp is more power efficient. The problem of the right half plane zero is also gone. Note that we ignored the current consumption of the current buffer in our calculations. This is because we will reuse a cascode transistor already present in the op-amp. To be totally correct, we should also include the input impedance of the current buffer in the calculations. This is done in [20].

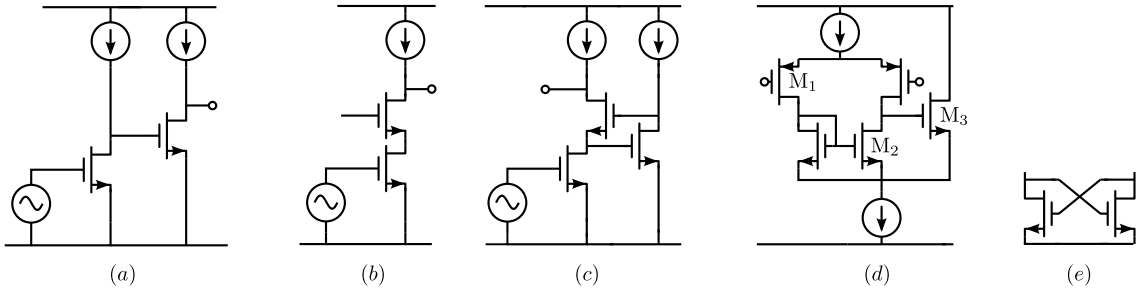


Figure 3.11: Gain enhancement techniques. (a) Cascade (b) Cascode (c) Gain boosting or regulated cascode (d) Bootstrapping (e) Current starving

3.6 Techniques to increase the DC gain

Since all the advantages (noise suppression, linearity,...) of a MOSFET-C integrator are due to the high gain of the op-amp, we want the op-amp used in the integrator to have very high DC gain. There are several ways to realise this:

Cascade More gain can be obtained by cascading gain stages. Since each stage introduces a dominant pole, the poles must be compensated when a cascade is introduced in an op-amp.

Cascode A cascode is a common-gate transistor, which acts as a current buffer. By inserting a cascode in a common-source amplifier, the output impedance increases and so does the DC gain of the amplifier. The Gain Bandwidth Product of the amplifier remains the same, only the DC gain increases. Drawback of cascodes is the lower output swing of the amplifier.

Gain boosting or regulated cascode By applying more feedback around a cascode, the gain can be even more increased without altering the GBW. An example of a Gain-Boosted cascode is shown in Figure 3.11c. The gain of the gain booster is added to the DC gain. The gain boosting stage is often realised with another op-amp, care has to be taken then to match the GBW of the gain booster to the GBW of original cascode ([5] pp. 82-84)

Bootstrapping Parasitic circuit elements can be cancelled out by placing the same voltage at both sides of it. In Figure 3.11d, source follower M_3 places the same AC voltage at both sides of transistor M_2, so the output resistance of M_3 is bootstrapped out so that the gain is determined by the $g_m r_o$ of transistor M_1 only.

Current-starving Positive feedback can create negative resistances. The cross-coupled pair shown in Figure 3.11e presents a negative resistance equal to $-2g_m$. By using a cross-coupled pair in combination with a normal load, it is possible to increase the total resistance seen at the nodes. Cancellation of the resistance by making all transistors the same size works only in theory, because mismatches can make the total structure unstable.

We shall use the first two techniques mentioned. Ahuja compensation calls for a combination of a cascade with a cascode anyway because we need a current buffer. If the DC gain found after the design is not sufficient, gain boosters can be added to the cascodes in order to obtain the right

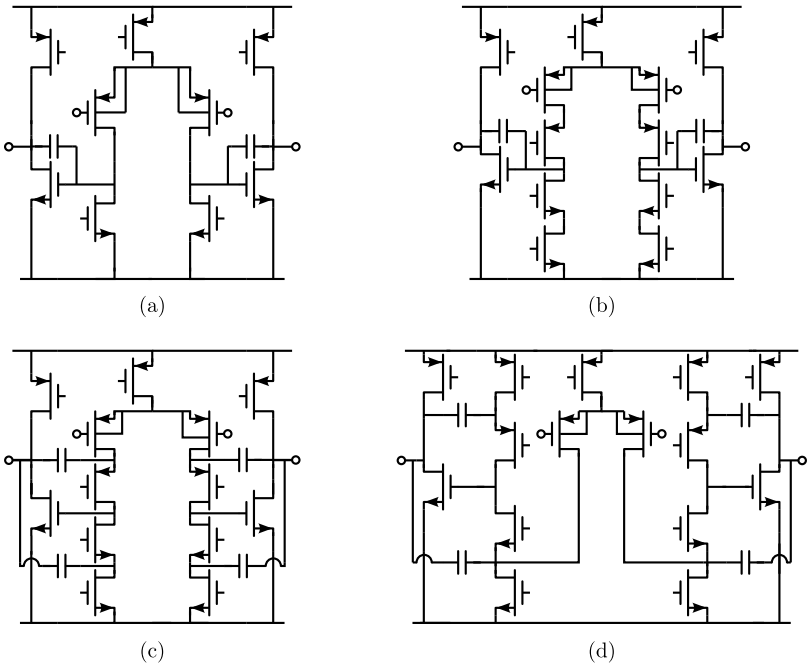


Figure 3.12: (a) Miller op-amp: A differential pair acts as the first stage, the second stage consists of two compensated common-source amplifiers. (b) telescopic cascode Miller-compensated op-amp: The first stage in the Miller op-amp is replaced by a telescopic cascode. (c) Ahuja compensated telescopic cascode op-amp: The cascodes in the previous op-amp are used as current buffers for the Ahuja compensation (d) Folded-cascode Ahuja compensated op-amp: The telescopic cascode from the previous op-amp is folded

gain. Bootstrapping creates difficult architectures and current starving can cause instability, so these techniques will not be used.

3.7 Integrator op-amp choice

Let's now choose an architecture for our op-amp. Using the know-how of the Miller and Ahuja compensation, we will build a two-stage op-amp. For the second stage, a common-source amplifier is chosen because it has a large gain, can drive a resistive load up to a certain degree and it has a large output swing. We want the input capacitance of the stage to be the as small as possible, so we pick an NMOS transistor for the signal with a PMOS load.

For the first stage we start with a normal differential pair. To increase linearity, we can connect the bulk of the input transistors to their source. To do this, we need either a triple-well NMOS or a PMOS transistor. Both are available in our technology of choice. We choose for a PMOS input transistor. The resulting fully differential op-amp is shown in Figure 3.12a. The gain of this op-amp will be in the order of $(g_m r_o)^2$ or about 40dB.

Since we want higher gain, we add cascodes to the input transistors and to the load transistors. The resulting op-amp is called the telescopic-cascode op-amp and is shown in Figure 3.12b. This op-amp will give a gain of about $(g_m r_o)^3$. The output swing of the first stage is drastically lowered because of

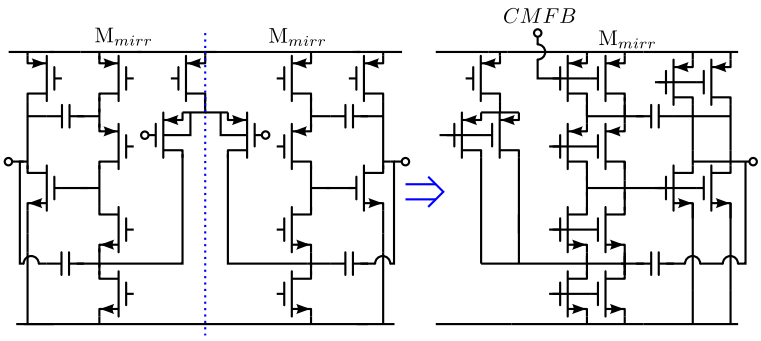


Figure 3.13: Looking at the op-amp from a common-mode view

the cascodes. In op-amp b, we are still using Miller compensation. However, since we have cascodes in the first stage, we can use the Ahuja compensation. We can choose between two cascodes to use as a current buffer. In [21] both of them are discussed and designed to get the best settling time. We shall choose a combination of both, with the ratio between them as another design parameter. The op-amp we have now is shown in figure 3.12c.

To satisfy the conditions for the Ahuja compensated op-amp, we need a large g_m in the current buffer. In the telescopic cascode op-amp, the current used for the buffer is equal to the current used for the input transistor, so the choice of g_m will be limited. In order to have freedom in the choice of g_m of the transistors forming the current buffers, we fold the cascode in the telescopic cascode op-amp. This results in the folded-cascode op-amp, shown in Figure 3.12d. This will be our op-amp of choice.

3.8 Common-mode feedback

In all the considered fully differential op-amps, the common-mode voltage of the first stage is not fixed because there are several competing current sources trying to set it. To fix this, a feedback mechanism is needed which corrects the common-mode output voltage. This correction can be performed at any of the 3 current sources of the first stage.

The design of the feedback loop for the common-mode is difficult, because the common-mode feedback loop already contains most of the poles and zeroes of the differential stage and the GBW of the common-mode loop should be equal or greater than the differential GBW. We start with the analysis of the common-mode viewpoint of the chosen amplifier and find the influence of the common-mode feedback limitations on the design of the differential amplifier. Second, we look into the different ways of measuring the common-mode signal and we choose a common-mode feedback configuration. Then it's time to start the sizing of our op-amp.

If we excite our chosen op-amp with a common-mode signal, we can use the symmetry in the architecture to assume the voltages left and right of the symmetry line are identical. We can thus assume these nodes are the same and actually short-circuit them. This results in the configuration shown in Figure 3.13. We will apply the common-mode feedback at transistor M$_{mirr}$. The dominant and nondominant pole frequencies of this circuit can easily be determined:

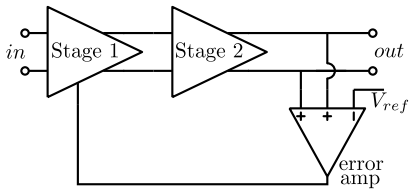


Figure 3.14: Model for the fully-differential op-amp with common-mode feedback

$$GBW_{cm} = \frac{g_{m,\mathrm{M}_{mirr}}}{2\pi C_C} \tag{3.24}$$

$$f_{nd,cm} = \frac{g_{m,\mathrm{M}_{out}}}{2\pi(C_L + C_C)} \frac{C_C}{C_{n,1}} \tag{3.25}$$

These expressions look similar to the ones of the differential system[4][19]. The nondominant pole lies at the same frequency as the nondominant pole of the differential circuit. This is because the second stage is the same for the differential-mode as for the common-mode. The GBW of the common-mode depends on $g_{m,\mathrm{M}_{mirr}}$, instead of $g_{m,\mathrm{M}_{in}}$ as was the case for the differential-mode.

During the design of the differential amplifier, the location of the nondominant pole is fixed to about 2 or 3 times the GBW (design parameter γ). The GBW of the common-mode should be the same or greater than the differential GBW to prevent instabilities on the common-mode inside the band of the op-amp.

The fact that the common-mode GBW must be equal or greater than the differential GBW demands that $g_{m,\mathrm{M}_{mirr}}$ must be greater or equal to $g_{m,\mathrm{M}_{in}}$. If a common-mode feedback amplifier is considered, the difference between $g_{m,\mathrm{M}_{mirr}}$ and $g_{m,\mathrm{M}_{in}}$can be compensated by adding gain to the amplifier. For example, if the GBW_{cm} is too small, it can be increased by placing some amplification in the error amplifier. It is clear that the difference in g_m of both transistors M_{mirr} and M_{in} will be small. Hence, the gain of the error amplifier will have to be close to unity.

The biggest problem for the common-mode loop is the location of the second pole. The common-mode loop should be stable, so a phase shift of more than 180° cannot be tolerated. The dominant pole is introduced by the first stage and will determine the GBW. The nondominant pole, determined by the second stage of the differential amplifier will introduce anther 90° phase shift and will kill the phase margin or even introduce instability. Because of this, the nondominant pole of the common-mode should lie at a higher frequency than the GBW of the common-mode.

This leaves little room for the GBW of the common-mode:

- the GBW_{cm} should be greater than the GBW_{dm}
- the GBW_{cm} should be smaller than $f_{nd,cm}$

To get a critically damped system in the differential amplifier, the nondominant pole was placed at twice the GBW. This leaves very little room for the GBW_{cm}, so we will place the nondominant pole further away.

[4]because all the elements in the common-mode system have a value which is the double of the value of the differential system, the factors 2 disappear in the fractions.

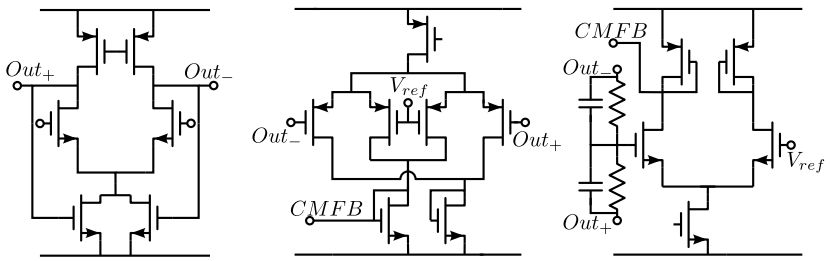


Figure 3.15: Common-mode feedback circuits. (a) linear region MOSFETS used as common-mode feedback elements (b) Full error amplifier (c) Error amplifier which uses resistors to measure the common-mode

To give us more design margin, we will place the nondominant pole further away from the GBW. This will give more room to place the GBW_{cm} and some mode phase margin, which we can use in the common-mode error amplifier.

Let's now look into the error amplifier. Its use is twofold:

1. Measure the common-mode at the output without influencing the differential-mode
2. Compare the measured common-mode to a wanted common-mode signal, amplify this error and feed it back into the amplifier.

Remember that the gain has to be limited because of the location of the poles. Because of this, diode connected transistors are commonly used as a load in these error amplifiers. There are several error amplifiers proposed in [5]

Linear region MOSFETS Figure 3.15a shows a differential pair with linear region MOSFETs as load. The value of these resistors can be varied by tuning their gate voltage. We can thus apply the common-mode feedback by tuning these gate voltages.

Error amplifier Using a special differential pair (shown in Figure 3.15b), we can calculate the difference between the common-mode presented at the output and a reference signal. Advantages of this technique are that the topology of differential circuit is not altered and that it only adds a capacitive load to the differential circuit. The biggest drawback is that the input range of this error amplifier circuit limits the differential signal swing of the differential stage.

Resistors and Error amplifier In order not to limit the differential swing, resistors can be used to determine the common-mode (Figure 3.15c). This way, only the common-mode output swing is limited by the error amplifier. To prevent the resistive loading of the differential stage, either the resistors are made very large, or source followers are inserted between the resistors and the output. When only large resistors are used, they tend to form a slow RC time constant with the input capacitance of the error amplifier and introduce a phase shift. This phase shift will have a negative impact on the phase margin of the common-mode feedback loop. To prevent this effect, capacitors can be inserted in parallel with the measuring resistors. This introduces an extra load for the differential stage though, so the capacitors should be small.

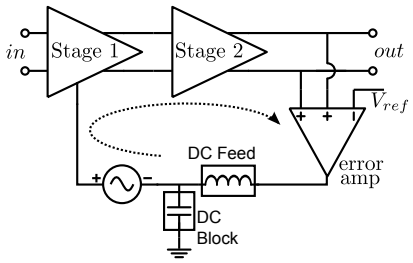


Figure 3.16: Opening the common-mode feedback loop in an AC simulation in ADS

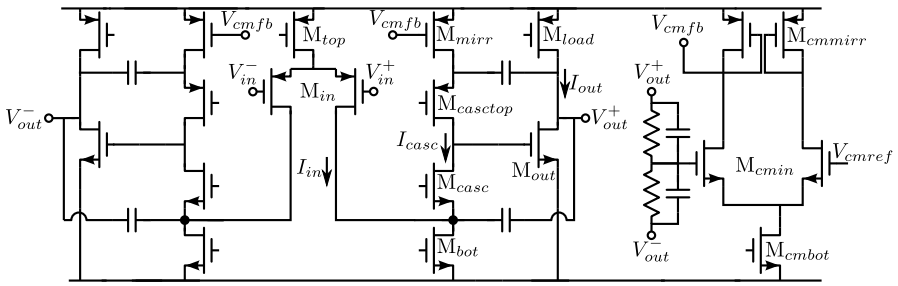


Figure 3.17: Fully differential folded-cascode op-amp with error amplifier

We shall use the last technique because of its obvious advantages. Since the op-amp has to be able to drive a resistive load anyway, adding the common-mode measurement resistors does not reduce the performance significantly.

In order to determine the loop gain, GBW_{cm} and phase margin of the common-mode loop in an AC simulation, we have to 'open the loop' [8]. In an AC analysis in ADS, this is done with two ideal components called DC FEED and DC BLOCK. They are used in the configuration shown in Figure 3.16. The DC FEED component blocks all AC signals, but the DC signals can pass through. This way, the DC operating point is not influenced by the opening of the loop.

The common-mode loop gain now corresponds to the AC FRF from one side of the DC FEED to the other side of it (indicated by the dashed line in the Figure). We need an AC source to determine this FRF. In order to ground this AC source without grounding the node at DC, the DC BLOCK component is used. It blocks all DC current, but is considered as a short for the AC signals.

With the described set-up, the common-mode loop gain is calculated and its stability is verified.

3.9 Design of the folded cascode op-amp

This section discusses the design of our fully differential folded-cascode op-amp. We want to be able to drive a load capacitance 10pF with a GBW of 100MHz. The definitions of the names of the transistors and currents are shown in Figure 3.17.

3.9.1 differential-mode

We start by choosing the design parameters:

$$\begin{aligned} \alpha &= \frac{C_L}{C_C} = 3 \\ \beta &= \frac{C_C}{C_{n,1}} = 3 \\ \gamma &= \frac{f_{nd}}{GBW} = 5 \end{aligned} \tag{3.26}$$

Note that γ is taken 5 to allow for more headroom in the design of the common-mode feedback circuit. Transconductance $g_{m,M_{in}}$ can be easily calculated by using the GBW, the parameter α and the load capacitance C_L.

$$g_{m,M_{in}} = \frac{2\pi GBW \cdot C_L}{\alpha} = 2.1\text{mS} \tag{3.27}$$

To counter $1/f$ noise and mismatches, we want transistor M_{in} to be large, so we choose a relatively low overdrive voltage of 0.12V which corresponds to a $^{g_m}/_{I_d}$ of $16,6$. With these choices, it is possible to size M_{in}

$$\frac{g_{m1}}{I_{in}} = 16.6 \quad \left(\frac{W}{L}\right)_{in} = \frac{2nI_{in}}{K_{p,PMOS}V_{ov}^2} = 241.6 \tag{3.28}$$

Since the current flowing into M_{in} is known, The transistor M_{top} can now easily be sized. We use an overdrive voltage of 0.2V and end up with a $^{W}/_{L}$ of 173.9.
The g_m of Transistor M_{out} is determined by the expression for the nondominant pole in the case of an Ahuja compensated op-amp and the factor γ:

$$f_{nd} = \gamma GBW = \frac{g_{m2}}{2\pi(C_C + C_L)}\frac{C_C}{C_{n1}} \tag{3.29}$$

Using the parameters α and β allows us to give an approximate expression for this equation and gives us a starting value for g_{m2}.

$$g_{m2} = 2\pi\gamma GBW \cdot C_L(1 + \frac{1}{\alpha})\frac{1}{\beta} = 14\text{mS} \tag{3.30}$$

We use the classical overdrive voltage of 0.2V or $^{g_m}/_{I_d} = 10$ for the output transistor. This allows us to size it:

$$\left(\frac{W}{L}\right)_{out} = \frac{2nI_{out}}{K_{p,NMOS}V_{ov}^2} = 351.8 \tag{3.31}$$

The PMOS load transistor of the output stage M_{load} can easily be sized with the current determined by M_{out}. With the same overdrive voltage, it will be about 3 times larger than M_{out}. The last unknown that remains is the current needed in the cascode stage. In the calculations about Ahuja compensation, we found that the g_m of the current buffer must be large to place the third pole introduced by the compensation at high frequencies. Since the common-mode feedback will be applied to M_{mirr} and since we want the common-mode to have the same GBW as the differential-mode, the g_m of M_{mirr} has to be close to the g_m of M_{in}.

Let's take the g_m of M_{mirr} the same as the g_m of the input transistor (namely 2.1mS), but with a normal overdrive voltage of 0.2V ($g_m/I_d = 10$). This results in the following sizing

$$\left(\frac{W}{L}\right)_{mirr} = \frac{nI_{casc}}{K_{p,PMOS}V_{ov}^2} = 144.9 \tag{3.32}$$

For the current mirror, we want a slightly larger g_m to make sure the pole is high frequent enough. To do this with a fixed bias current, we lower the overdrive voltage to 0.15V. This will increase the size of this transistor compared to M_{mirr}

$$\left(\frac{W}{L}\right)_{casctop} = 257.7 \tag{3.33}$$

For the NMOS cascode M_{casc} we apply choose the same overdrive voltage and bias current as with $M_{casctop}$. Since it's an NMOS, it will be about 3 times smaller

$$\left(\frac{W}{L}\right)_{casc} = 93.8 \tag{3.34}$$

The last transistor that remains for the differential amplifier is the bottom transistor. Its bias current is the sum of the input current and the cascode current. The overdrive voltage is again chosen to be 0.2V.

$$\left(\frac{W}{L}\right)_{bot} = \frac{n(I_{casc} + I_{in})}{K_{p,NMOS}V_{ov}^2} = 84.4 \tag{3.35}$$

3.9.2 Common-mode

The design of the error amplifier uses a lot of elements of the differential amplifier to re-use bias voltages of the differential amplifier. In order to get a gain close to unity, a diode connected load is used for the error amplifier. Its gain is $g_{m,cmin}/g_{m,cmload}$. We want it to be unity, so $g_{m,mcin}$ must be equal to $g_{m,cmload}$.

We want the GBW of the error amplifier to a few times larger than the GBW of the differential-mode. The ratio between the GBW of the differential-mode amplifier and the error amplifier will be called δ.

$$\delta = \frac{GBW_{error}}{GBW_{diff}} \tag{3.36}$$

In order not to decrease the phase margin determined by the poles of the differential amplifier, we choose a δ of 5.

The load capacitance of the error amplifier is twice the gate capacitance of M_{mirr}. We assume it to be 1pF.

$$g_{m,cmin} = \delta C_{GS,mirr} 2\pi GBW \tag{3.37}$$

We want to reuse the load transistor M_{cmmirr} of the error amplifier to provide a bias value for M_{mirr}, so the overdrive value of that transistor is fixed. Since we want both transistors to have the same g_m

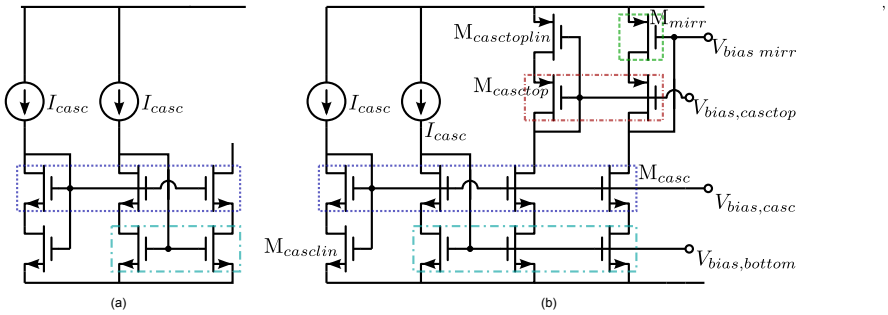


Figure 3.18: Bias Circuits for the cascode. Matched transistors are encircled.

and since both transistors use the same bias current, the overdrive voltages of transistors M_{cmin} and M_{cmmirr} must be the same. With this information, the transistors can be sized

$$\left(\frac{W}{L}\right)_{cmin} = 79.1 \quad \left(\frac{W}{L}\right)_{cmload} = 217.4 \quad \left(\frac{W}{L}\right)_{cmbot} = 158.3 \tag{3.38}$$

The value of the resistors used to determine the common-mode has to be designed too. These resistors form a low-pass filter with the gate of M_{cmin}. Ideally, the resistance value should be very high. However, for a fixed capacitance, the frequency of the introduced pole decreased. The pole of this low-pass filter introduces a phase shift which kills the common-mode phase margin. Therefore, we place a capacitor in series with the resistor so that the pole is compensated with a zero.

3.9.3 Bias circuit

A commonly used bias circuit for a cascode current source is shown in Figure 3.18a. It uses a copy of the cascode in a current mirror. The bias voltage of the cascode itself is still not known. It can be obtained using a simple current mirror, but the operating conditions of that current mirror don't resemble the one for the cascode. Hence a bad result will be obtained. The solution to this problem is shown in Figure 3.18a. The bias voltage for the cascode is obtained with two transistors: the top of those is an exact copy of the cascode transistor in the current mirror. The bottom transistor will always be in it's linear region. We will use it to provide the right voltage at the source of the cascode to set the cascode transistor in the right operating conditions. Let's repeat the expression for the current in a triode region MOSFET:

$$I_{ds} = \frac{K_P}{n}\frac{W}{L}\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right)(V_{DS}) \tag{3.39}$$

We want the operating point of the top transistor to be the same as the one in the cascode. The V_{DS} is thus determined by the overdrive voltage of the bottom transistor with an added margin (V_{margin}) to ensure the bottom transistor remains in saturation when process variations come into play. The V_{GS} of the transistor should be the V_{GS} of the cascode transistor, so it is the sum of the overdrive voltage of the bottom transistor $V_{ov,bot}$, the margin V_{margin}, a V_T and the overdrive voltage of the cascode $V_{ov,casc}$. We find

$$I_{ds} = \frac{K_P}{n}\frac{W}{L}\left(V_{ov,bot} + V_{margin} + V_T + V_{ov,casc} - V_T - \frac{1}{2}\left(V_{ov,bot} + V_{margin}\right)\right)\left(V_{ov,bot} + V_{margin}\right)$$

$$I_{ds} = \frac{K_P}{n}\frac{W}{L}\left(\frac{1}{2}V_{ov,bot} + \frac{1}{2}V_{margin} + V_{ov,casc}\right)\left(V_{ov,bot} + V_{margin}\right) \tag{3.40}$$

This expression can be used to size the linear region bias transistor $\mathrm{M}_{cascbias}$. The bias circuit in Figure 3.18a can also be used to obtain the biasing for the PMOS part of the cascode stage. This is done by just taking the PMOS equivalent of the circuit shown in Figure 3.18a. The current for the PMOS bias circuit can be obtained by placing a biased cascode current source, as shown in Figure 3.18b. The circuit shown there is the total resulting bias circuit for the cascode.

To lower the current consumption, the W/L of all the transistors in the bias circuit can now be scaled. To calculate the maximal allowed scale factor, one should take mismatches into consideration. Since the current consumption is not critical for our chip, we don't scale the transistors of the bias circuit[5].

Applying (3.40) for the sizing of the linear transistors in the bias circuit, we obtain

$$\left(W/L\right)_{\mathrm{M}_{casclin}} = 15.3 \qquad \left(W/L\right)_{\mathrm{M}_{casctoplin}} = 42.1 \tag{3.41}$$

The current going through the bottom transistor in the bias circuit is I_{casc}, as shown in the figure. In the op-amp, the DC current flowing through M_{bot} is $I_{in} + I_{casc}$. If we would just copy M_{bot} without changing its W/L, the biasing obtained by just sending I_{casc} trough the transistor would be wrong. Therefore, we rescale the W/L of the bottom transistor used in the bias circuit (called M'_{bot} from now on)

$$\left(W/L\right)_{\mathrm{M}'_{bot}} = \frac{nI_{casc}}{K_{p,NMOS}V^2_{ov,\mathrm{M}_{bot}}} = 105 \tag{3.42}$$

In order to have maximum swing available at the output of the first stage, we want the DC operating point of the output of the second stage to be at half the supply voltage, or 0.9V. If we do this, we fix the DC V_{GS} of the input transistor of the second stage and hence its operating point. Because V_T of the used transistors lies around 0.4V, the overdrive of the output transistor would have to be 0.5V to obtain 0.9V in total. This makes the transistor inefficient and limits the output swing.

The process used also provided transistors for a supply voltage of 3.3V. Their V_T lies around 0.7V. If we choose one of these transistors, a better overdrive voltage of 0.2V yields the wanted 0.9V DC operating voltage at the output of the first stage. Because of this, the transistor M_{out} is a $3.3V$ transistor. The process parameters were assumed to be the same for the 3.3V transistors as for the 1.8V transistors, so the sizing of M_{out} was not changed.

[5]The proposed bias circuit provides all needed bias voltages for the cascode stage of the folded cascode. One of them is set by the common-mode feedback circuit, so we don't actually need to place the rightmost branch of the bias circuit.

3.10 Simulation results

Now that the $^W/_L$ and the overdrive voltage of all the transistors are chosen, the gate-lengths are the only parameter left to determine. Like with the Miller op-amp from the design example earlier, we will do this by tuning the values in ADS. The transistors are divided into four groups with matched gate-lengths:

1. L_{in}: Input transistors M_{in} and the tail transistor M_{top}.
2. L_{out}: The transistors of the output stage M_{out} and M_{load}.
3. L_{casc}: The transistors in the cascode stage M_{bot}, M_{casc}, M_{mirr}, $\mathrm{M}_{casctop}$ and their biasing transistors. Also the load transistor in the common-mode feedback stage[6] is matched to this gate-length.
4. L_{cm}: The input transistors of the common-mode feedback stage M_{cmin} and the tail transistor of the common-mode feedback stage M_{cmbot}

Tuning was started with all gate-lengths equal to 1μm. At this gate-length, all specifications were reached, so the tuning didn't have to be performed. The results of the AC simulations are shown in Figures 3.19 and 3.20. The gain of the common-mode feedback stage is not unity as shown in Figure 3.23. Because of this, the GBW of the common-mode feedback loop is lower than the differential GBW. This can be solved in two ways:

- Increase the gain of the error amplifier and make it unity.
- Increase $g_{m,\mathrm{M}_{mirr}}$. Looking at (3.24), this will increase the GBW of the response from common-mode feedback to the output of stage 1.

Doubling $g_{m,\mathrm{M}_{mirr}}$ it solved the problem, making the GBW_{cm} equal to GBW_{dm}. This action increased the current though the cascode stage, so the sizing of the transistors M_{bot}, M_{casc}, M_{mirr} and $\mathrm{M}_{casctop}$ was changed. Their new $^W/_L$ was calculated with the quadratic model as shown before. The final values for all transistors are shown in Figure 3.25. The results of AC simulations performed on this final design are shown in Figures 3.21 and 3.22. The phase margin of the common-mode feedback loop is 28°, which is low, but within acceptable bounds. The step response of the op-amp in unity feedback is shown in Figure 3.24.

The op-amp has been designed and it satisfies the specifications. A DC gain of over 100dB is obtained with a GBW of 74MHz. A lot of specifications have not been checked though. A thorough noise analysis should be performed. The power supply rejection ratio should be estimated and mismatches should be incorporated into the design. The effect of process variations should be analysed and a lay-out should be made before the op-amp can be used in the chip. We chose to to take another route instead.Because the main performance killer of the MOSFET-C filter is nonlinear distortion, we will look into the nonlinear behaviour of the op-amp. In the next chapter, an analysis method is developed for the op-amp which could be used in a general optimisation of the op-amp and the complete MOSFET-C filter. Applying this method and actually building the chip are left as future work.

[6] This is because the DC biasing of the transistor M_{mirr} is determined by the load transistor of the common-mode feedback circuit

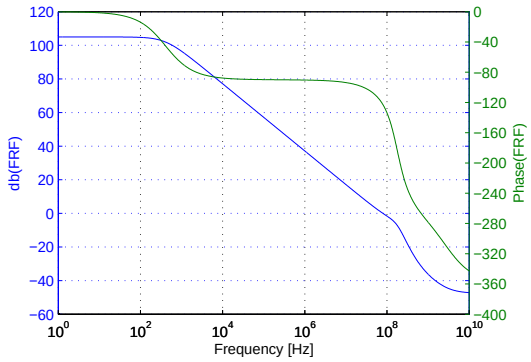


Figure 3.19: Differential open loop gain of the op-amp obtained with the design strategy. (-) is the amplitude in dB. (-) is the phase in degrees. The GBW is 80MHz and the phase margin is 58°.

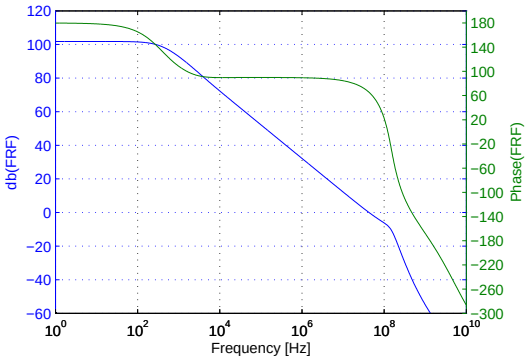


Figure 3.20: Loop-gain of the common-mode feedback loop of the op-amp obtained with the design strategy. (-) is the amplitude in dB. (-) is the phase in degrees. The GBW is 42MHz and the phase margin is 65°.

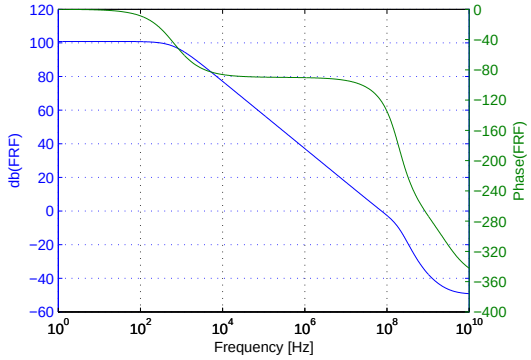


Figure 3.21: Differential open loop gain of the corrected op-amp. (-) is the amplitude in dB. (-) is the phase in degrees. The GBW is 74MHz and the phase margin is 58°.

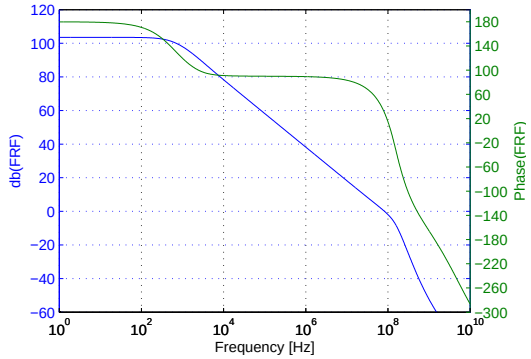


Figure 3.22: Loop gain of the common-mode feedback loop of the corrected op-amp. (-) is the amplitude in dB. (-) is the phase in degrees. The GBW is $83MHz$ and the phase margin is 28°.

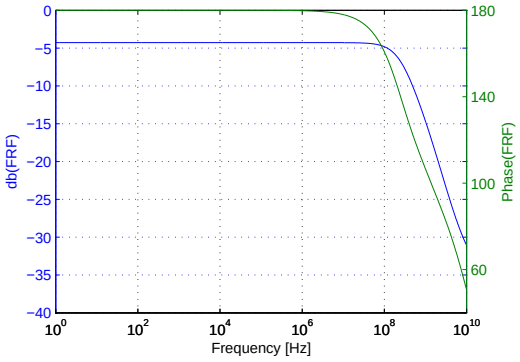


Figure 3.23: FRF of the error amplifier. (-) is the amplitude in dB. (-) is the phase in degrees.

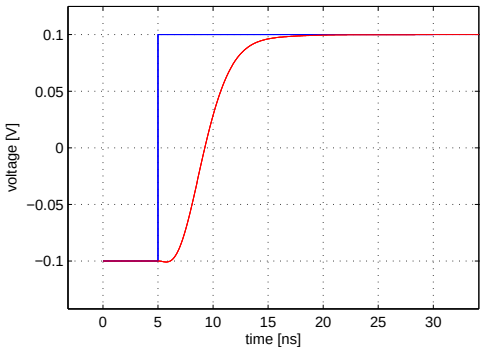


Figure 3.24: Step response of the op-amp when placed in a unity feedback configuration

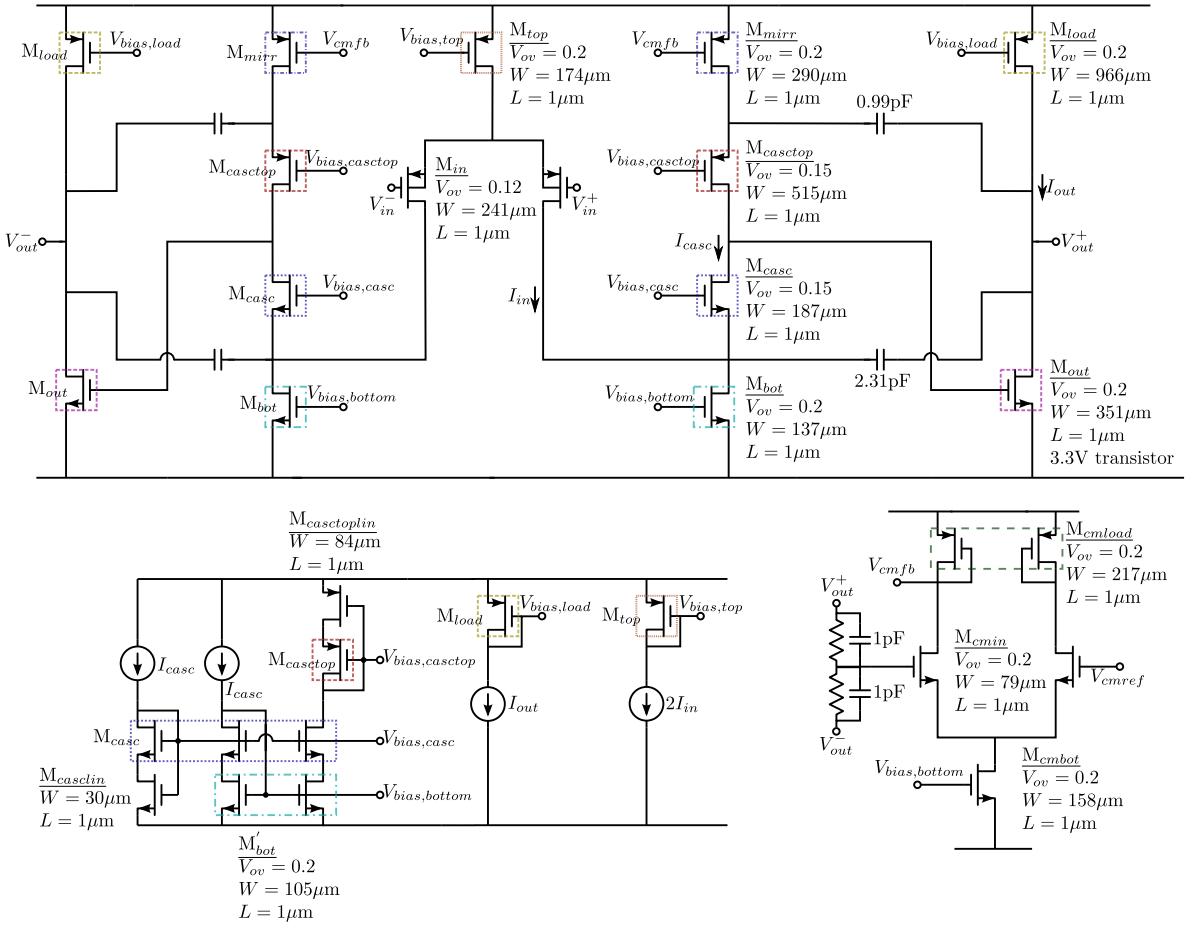


Figure 3.25: Final design of the Ahuja compensated folded-cascode op-amp

Chapter 4

Nonlinear Analysis of the Op-amp

The previous chapter discusses the design of two-stage op-amps. In order to improve it, we will develop a simulation-based method to determine which of the stages in the op-amp contributes the most to the nonlinear distortion at the output when placed in a feedback configuration. The result of this analysis can be used to improve the performance of the responsible stage.

A conference paper was written about this analysis method and accepted for the International Conference on Synthesis, Modelling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)[1]. Therefore, the paper is presented in its original lay-out.

There are some major differences between the op-amp used in the analysis method described in the paper and the op-amp designed for the programmable filter:

1. The simulation-based method is developed on a single-ended op-amp because the extra feedback loop for common-mode stabilisation of a fully differential op-amp adds another difficulty to the analysis.

2. The op-amp in the analysis is a Miller compensated op-amp. In an Ahuja compensated op-amp, where one or more of the cascodes of the first stage are used as current buffer, the first and second stage are mixed up. Because of this, it is not possible to make a clear distinction between the first and the second stage of the op-amp.

In this chapter, after presenting the paper, we will look more closely into the errors made in the analysis method by considering the difference between the AC simulations used in the analysis an the BLA of the stages. Afterwards, the method is extended and used to analyse a fully-differential Miller-compensated op-amp.

[1] www.smacd2012.org

Determining the Dominant Nonlinear Contributions in a multistage Op-amp in a Feedback Configuration

Adam Cooman, Egon Geerardyn, Gerd Vandersteen, Yves Rolain

***Abstract*—In this paper a simulation based method is proposed to determine the position of the dominant nonlinear contribution in the schematic of multistage op-amp operated in a feedback configuration. The key idea is to combine the Best Linear Approximation (BLA) and a classical noise analysis to determine the dominant source of nonlinear contributions. This results in a powerful yet simple design tool which does not require special analyses or custom models. As an example, the method is applied to a folded-cascode op-amp.**

***Index Terms*—nonlinear distortion, operational amplifiers**

I. Introduction

MOST analog design flows rely only on linear time invariant reasoning while designing an analog/RF circuit. When the linear design flow is completed, the importance of nonlinearities is assessed by identifying compression points or intercept points. Those provide a measure of the nonlinear behavior of the total circuit only. This standard approach does not give any clue to help the designer to modify the design to decrease the nonlinearities as no information is provided about the source of the nonlinear distortion.

In [1]–[3] a Volterra-based approach was used to localize the nonlinearitiy of the circuit in an analytic way. For larger circuits, this analytic method yields lengthy complex expressions. Overview is hence easily lost. Those methods also require the replacement of the transistor model by an approximate analytic-nonlinear model.

In this paper, a method is proposed which can be positioned in between the linear design framework and the symbolic Volterra theory. The op-amp is considered to consist of a cascade of two or more gain stages. Each stage is considered as a black box. No knowledge about the interior of a stage is used. The nonlinear distortion is determined by a transient analysis. The input and output signal of every stage is measured during the simulation. The only constraint imposed on the transient simulation is the choice of the excitation signal used. A so called multisine excitation allows one to determine the BLA of the system. The BLA consists of an FRF model and a colored power spectral noise source to model the influence of the nonlinearity [5]. One can hence consider the nonlinear distortions as an additional colored noise source. Using regular noise analysis now allows one to determine the distortion that is introduced by every stage.

Section II explains the method and the theory behind it in more detail. Then, in Section III, the method is applied to a folded-cascode op-amp.

II. Methodology

In this section we explain the theory behind the method. First we define the multisine excitation signal that is used in the transient analysis. Using a special multisine allows to split the even and odd nonlinear contributions. They can be "measured" separately. Second, the theory of the BLA is introduced. This leads to the description of the nonlinear contributions as a colored Gaussian noise source. Finally we apply the noise analysis on the cascaded stages.

A. Multisine excitation

In [4] an odd random-phase multisine is shown to be a well-suited excitation signal for the detection of nonlinearities in a measurement context. This claim remains valid for simulations. Random-phase multisines combine the best of random excitations and periodic signals.

- Random excitations are close to real world signals. They allow a broad measurement bandwidth at the cost of spectral leakage and a reduced signal-to-noise ratio at some frequencies. They also hamper an easy detection of the nonlinearity.
- Periodic signals have a deterministic spectrum. They do not mimic real world signals very well. When properly designed, they don't suffer from spectral leakage and can ease the detection of the nonlinearity.

A random phase multisine behaves as a random noise signal that mimics real world signals but comes with the high signal-to-noise ratio and the nonlinear detection capability of a periodic signal. A random-phase multisine with N components is described by:

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} A_k \cos\left(2\pi k f_0 t + \phi_k\right) \tag{1}$$

where A_k and ϕ_k are the amplitude and phase of the k^{th} spectral line and f_0 is the resolution frequency of the multisine. The value of the phase spectrum is the result of a uniform random process over $[0, 2\pi[$. By imposing additional constraints on the frequency grid, it is possible to construct a multisine suited for the analysis of nonlinearities: the odd random phase multisine [5].

In this signal, only odd frequency bins are present ($A_{2k} = 0$). An even nonlinearity produces components at a frequency which is the sum of an even number of excited frequencies. As the excited frequencies are all chosen to lie on an odd frequency grid, the sum of an even number of such frequencies will lie on an even grid. This means that even nonlinearities

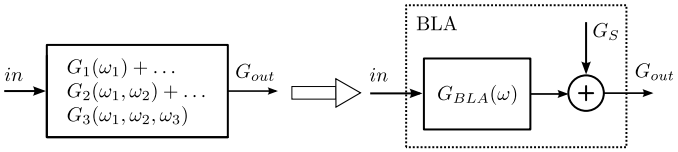


Figure 1. Applying the BLA on a nonlinear stage

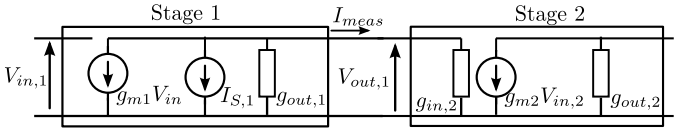


Figure 2. Cascade of two stages with a finite input impedance

will not interfere with the response of the system at the excited frequencies. Hence an even distortion becomes measurable by looking at the spectrum present at the even frequency grid lines.

The odd order nonlinear contributions will always contribute to the excited frequency lines. The sum of an odd number of odd frequencies always yields an odd frequency.

Two approaches exist to determine the level of odd non-linearities [6]. In this paper the faster of both is used. Some odd frequencies are omitted in the excitation signal. One then interpolates the measured distortion spectra at the non-excited frequency lines (also called the "detection lines") to estimate the odd non-linear contribution present at the excited lines [6]. To choose which excitation frequencies to omit, the excited bins are grouped into groups of 4 neighboring excited bins from which one bin is randomly removed.

B. Best Linear Approximation

Linear system theory describes the response of an LTI system as

$$Y(f) = G_0(f) \cdot U(f) \tag{2}$$

with $U(f)$ and $Y(f)$ respectively the deterministic linear input and output spectra and $G_0(f)$ the frequency response function (FRF) of the system. For nonlinear systems, this relation is no longer generally valid, but can be used to approximate the linearized behavior of the system around an operating point in least squares sense. This approximation is called the Best Linear Approximation (BLA). For a random multisine excitation with a sufficiently large number F of excited frequencies, the FRF can be written in the form [6]

$$G(f) = G_{BLA}(f) + G_S(f) + G_N(f) \tag{3}$$

with:

- $G_{BLA}(f)$ the best linear approximation. It consists of the linear term $G_0(f)$ and a systematic nonlinear bias term $G_B(f)$ which describes the compression/expansion of the system and is caused by odd nonlinearities,
- $G_S(f)$ the stochastic nonlinear contribution which acts as a noise source with zero mean,
- $G_N(f)$ the simulation (or measurement) noise.

This BLA represents the response of the nonlinear system to signals with similar properties (e.g. same power spectrum, probability density function, ...) as the signal applied to determine the BLA.

The BLA assumes that the system consists of an FRF $G_{BLA}(f)$ with an additive output noise source G_S which accounts for the stochastic nonlinear contributions (see Figure 1). Since the stochastic nonlinear contributions act like noise, it is possible to apply techniques borrowed from classical noise analysis on these nonlinear contributions. When we apply the BLA to every stage of the op-amp, we get a (nonlinear) noise source $G_S(f)$ for every stage. If we refer all the (nonlinear) noise sources in the system to the output, we can compare their contribution to the total measured output distortion of the system.

C. Determining and comparing the nonlinear contributions of each stage in an Op-amp

In order to refer the (nonlinear) noise contribution of each noise source to the output node, we need to know:

1) the power spectral density (PSD) of each noise source
2) the FRF between that noise source and the output node

If the stages behave dominantly linear, it is possible to perform the noise analysis using AC FRFs only. From now on, we will neglect the nonlinear bias term $G_B(f)$. To verify the validity of this assumption, it is sufficient to compare the (noisy) G_{BLA} obtained by the division of the spectra calculated during the transient analysis to the noise-free AC FRF G_0. The trade-off to be made is a classical bias versus variance trade-off. We have chosen to allow for a bias of a few dB in the amplitude and a few degrees in the phase in this paper.

1) Determining the PSD of the noise source: ideal case: To determine the PSD of the (nonlinear) noise source for a certain stage, we calculate the difference between the simulated nonlinear response of that stage and its linearized response. Looking at Figure 1 we find that in general:

$$G_S = G_{out} - G_{BLA} \cdot in \tag{4}$$

If we neglect the nonlinear bias term G_B, G_{BLA} boils down to the AC FRF G_0. If the input and output loading impedance of the stage are infinite, we can use the voltages measured at the input and the output port alone. This results in the ideal behavior

$$G_{S,i} = V_{out,i} - \mathrm{G}_{0,i} \cdot V_{in,i} \tag{5}$$

where $G_{S,i}$ is the (nonlinear) noise contribution, $V_{in,i}$ and $V_{out,i}$ are the voltages measured at the input and output of the i^{th} stage during the transient analysis respectively and $G_{0,i}$ is the FRF of the i^{th} stage, calculated with an AC analysis.

2) Determining the PSD of the noise source: real world case: In an op-amp, it's not possible to consider the input impedance of the stage to be infinite. Expression 5 will therefore yield a poor approximation of the PSD in this case. When the input impedance of the loading stage is not infinite, theory requires us to apply a full two-port noise analysis. To avoid the complexity, we have chosen an intermediate solution.

We use the Norton equivalent for the output of the stage-under-test and consider the (nonlinear) noise source to be a current source only. This neglects the voltage noise source that is present in the full 2-port case [7].

Consider a cascade of two stages as shown in Figure 2. Determining the noise contribution $I_{s,1}$ can be done using the following formula:

$$I_{s,1} = \frac{g_{in,2} + g_{out,1}}{g_{in,2}} \left(I_{meas} - \mathrm{FRF}_{V_{in,1} \to I_{meas}} V_{in,1} \right) \quad (6)$$

where $g_{in,2}$ is the input conductance of stage 2 and $g_{out,1}$ is the output conductance of stage 1. I_{meas} is the current flowing out of stage 1 and into stage 2 and $V_{in,1}$ is the input voltage of stage 1. Both are measured during the transient analysis. $\mathrm{FRF}_{V_{in,1} \to I_{meas}}$ is the transconductance of the first stage. All conductances and the transconductance are determined using an AC analysis.

3) Determining the FRF between the source and the output: The FRF needed to refer the calculated noise contribution to the output is determined using another AC analysis. An AC source is placed at the assumed location of the (nonlinear) noise source and it's response to the output is calculated.

- For nonlinear contributions calculated using (5), an AC voltage source is placed in series with the considered stage.
- For nonlinear contributions calculated using (6), an AC current source is placed between the output of the considered stage and AC ground.

D. Simulations

The necessary simulations were performed using classical AC and transient analysis, while the post-processing of the data was done in MATLAB.

First, the multisine excitation signal is generated in MATLAB and then imported into the transient simulation as a time-domain waveform. The sampling frequency of the simulation is chosen to be 10 times the maximum frequency of the multisine.

For an n-stage op-amp, the following simulations are needed:

1) A transient analysis with a multisine excitation to determine the nonlinear contributions at the unexcited frequency bins of the multisine. The op-amp can be placed in an inverting feedback configuration.
2) One AC analysis to determine the FRF of the stages. A voltage to current FRF is needed for the stages which are followed by a stage with a finite input impedance. A voltage to voltage FRF is obtained for the stages with an infinite load. This analysis can also be used to determine the input impedance of the stages, by measuring the current flowing into the stage.
3) For each stage, an AC analysis is needed to determine the total conductance of it's output node. For this AC analysis, the input of the stage is AC grounded.
4) For each stage, an AC analysis to determine the FRF from the considered noise source to the output. An AC source is added at the location of the equivalent nonlinear noise sources and it's influence is measured at the output.

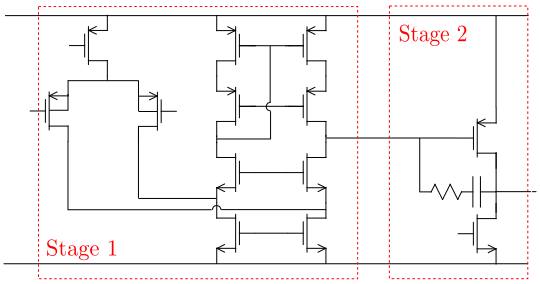


Figure 3. Op-amp used for the simulations

The set of AC analyses is not only used to determine the FRF of the subsystems, but also their input and output impedance and the impact of the different nonlinear contributions to the output. The latter enables the use of this nonlinear analysis in a classical noise analysis. The relative importance of each nonlinear source is obtained assessing its relative contribution to the total nonlinear distortion at the output. This results in an easy to use analysis tool to determine the dominant sources of nonlinear distortion.

The AC analyses are performed up to the sample frequency of the analysis, with a resolution determined by the lowest frequency of the multisine.

For the transient analysis, a fixed time step is chosen in function of the sampling frequency. Two periods of the multisine are simulated. The first period is discarded to suppress transient effects. The integration method is trapezoidal to prevent artificial damping of the poles in the op-amp, such that the results of the transient analysis match the results of the AC analysis up to the frequency where warping starts to occur [8].

III. Example: Folded-Cascode Op-amp

As an example, the developed method is applied to the folded-cascode op-amp shown in figure 3. The op-amp is designed for the UMC.18 CMOS technology. During the simulations, a BSIM3v3 model is used for the MOSFETs. The op-amp under test has a gain bandwidth product of 100 MHz and a DC gain of 80 dB. It is connected as an inverting amplifier with a gain of 10. The impact of the resistive loading of the output stage is reduced by a voltage buffer inserted between the output and the feedback resistor.

The multisine used for the experiment has a resolution of 100 Hz and excites frequencies up to 10 MHz. The sample frequency of the simulation is 100 MHz. The phase spectrum of the multisine is random. Note that a set of 100 realizations of the multisine was used to select the signal with the smallest crest factor. Its amplitude is scaled such that the output covers 80% of the supply voltage. This prevents clipping and imposes that the stages behave dominantly linear. The linear FRF, determined with the AC analysis can therefore be used.

The following results will be discussed: first, we analyze which stage contributes most to the nonlinear distortion. Next,

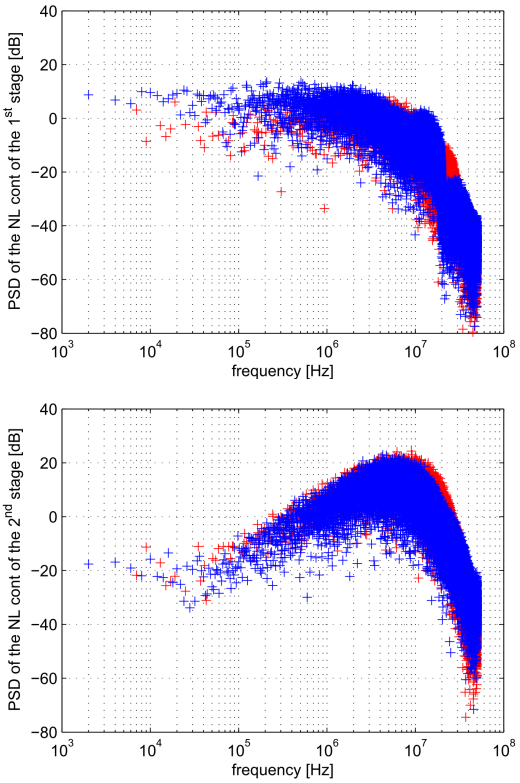


Figure 4. Output referred nonlinear contributions of the first and second stage. (+) are the contributions at the even frequency bins and (+) are the contributions at the odd frequency bins.

it is shown that the total distortion is equivalent to the sum of all distortion contributions.

The calculated nonlinear output contributions of both stages are shown in Figure 4. Blue symbols show the contributions at the even spectral lines, representing the even nonlinear distortion. Red symbols represent the contribution at the odd spectral lines without excitation, representing the odd nonlinear contributions. At low frequencies, the first stage is the dominant source of nonlinear distortion. At frequencies close to the gain bandwidth product, the second stage is responsible for most of the distortion.

To verify whether the contributions are correct, the sum of both output referred contributions is compared to the actual measured output spectrum during the transient simulation. The result of this comparison can be seen in Figure 5. The difference between the sum of the calculated contributions and the measured distortion at the output gives a measure for the error level of the procedure. Because the sample frequency of the transient analysis is 100 MHz, the results can only be considered to be accurate in a frequency up to about 10 MHz. The simulations show that the error increases at very low frequencies. This is due to the fact that the measured current between the stages is used to calculate the contribution of the first stage. At very low frequencies, the input impedance of the second stage is very large. Hence the current becomes very small, the numerical precision of the calculations comes into play and this increases the error.

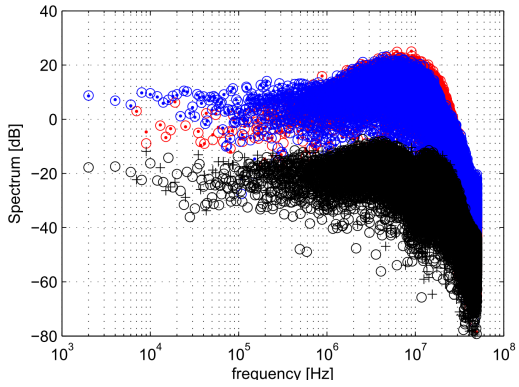


Figure 5. Comparison between the sum of the calculated output referred nonlinear contributions and the actual output spectrum measured during the transient simulation. (o) and (o) represent the measured output distortion at odd and even frequency bins respectively. (·) and (·) represent the sum of the calculated nonlinear contributions of both stages at odd and even bins respectively. (+) and (o) represent the difference between both at odd and even bins respectively.

IV. Conclusions

A transient simulation using a multisine excitation allows the extraction of a "best" linear transfer function and an equivalent nonlinear "noise" source. If the system behaves dominantly linear, one can use the AC analyses to determine the output distortion generated by each stage. Assuming the nonlinear distortion behaves as a current noise source allows to take finite input impedance of the stages into consideration without using a full two-port noise analysis. The calculated distortion is then referred to the output by simulating the AC transfer function between the assumed source and the output. By doing so, the nonlinear contribution of each stage in an op-amp to the output is determined. This method allows to determine the dominant source of nonlinearities without using special simulation techniques or models.

References

[1] P. Wambacq and W. Sansen, "*Distortion Analysis of Analog Integrated Circuits,*" Norwell, MA: Kluwer, 1998.

[2] B. Hernes and W. Sansen, "Distortion in Single-, Two- and Three-Stage Amplifiers," *IEEE trans. Circuits Systems*, Vol. 52 No. 5, May 2005

[3] S. O. Cannizzaro, G. Palumbo and S. Pennisi, "Distortion Analysis of Miller-Compensated Three-Stage Amplifiers," *IEEE trans. Circuits Systems*, Vol. 53, No. 5, May 2006

[4] J. Schoukens, J. Swevers, R. Pintelon, H. Van der Auweraer, "Excitation design for FRF measurements in the presence of nonlinear distortions," *ISMA 2002, Noise and Vibration Engineering*, Leuven, 16-18 September 2002, pp. 951-958

[5] R. Pintelon and J. Schoukens, "*System Identification, A Frequency Domain Approach,*" IEEE Press, 2001.

[6] R. Pintelon, G. Vandersteen, L. De Locht, Y. Rolain and J. Schoukens "Experimental Characterization of Operational Amplifiers: a System Identification Approach—Part I: Theory and Simulations," *IEEE Trans. on Instrumentation and Measurement*, Vol. 53, No. 3, June 2004, pp. 854-862

[7] J. Engberg and T. Larsen, "*Noise Theory of Linear and Nonlinear Circuits,*" Wiley, 1995

[8] K.S. Kundert, "*The Designer's Guide to SPICE and Spectre,*" Springer, 1995

Before we go into more detail in some of the methods explained in the paper, let's repeat the basics of the analysis method:

1. The BLA allows us to consider nonlinear distortion as noise.
2. Using a random odd random phase multisine (A random-phase multisine which excites only odd frequency bins of which some odd excitation lines are omitted), we can split the nonlinear contributions generated by even-order nonlinearities from the ones generated by odd-order nonlinearities.
3. Applying spectral correction[2] to the output of a page, we can calculate the nonlinear contribution of each stage at its output.
4. Using the FRF of each nonlinear noise source to the output, we can refer all the sources to the output and compare their influence to the total nonlinear distortion generated by the op-amp.

4.1 BLA versus AC

Theoretically, the BLA of the stages should be used in the spectral correction of the second step. Using the definitions found in Figure 4.1, we find

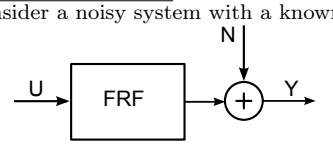
$$G_S = G_{out} - G_{BLA} \cdot in \tag{4.1}$$

where G_S is the nonlinear contribution of the stage under consideration, G_{out} the output signal, G_{in} the input signal and G_{BLA} is the BLA of the stage under consideration. In the paper, the AC FRF of the stage was used instead of the actual BLA. This has two good reasons:

1. In the considered feedback configuration, it would be very hard to determine the BLA. Due to the large gain of the op-amp at low frequencies, the input signal at the op-amp input is very small (See Figure 4.3). Hence, the resulting BLA is very noisy at low frequencies.
2. Even at high frequencies, the BLA obtained using the fast method is noisy. To counter this, averaging over several realisations of the multisine should be applied. The transient simulation is already the most time-consuming simulation, so having to perform several transient simulations would increase the simulation time to unacceptable levels.

The AC FRF of the stages has the advantage that it is noiseless and can be determined very quickly[3].

[2] Spectral correction:
Consider a noisy system with a known FRF.



If the input and output spectra are exactly known, we can calculate the noise generated by the stage by subtracting the known response to the input from the output: $N = Y - FRF \cdot U$

[3] In the currently implemented version of the analysis, all points of the multisine are also calculated using an AC analysis. Due to this, the simulation time of the AC analysis is also quite long. It can easily be decreased though: by making the frequency grid for the simulation coarser and interpolating its result to all the frequency bins of the multisine. When this interpolation is applied, the transient simulation will definitely be the most time consuming of both simulations.

There will be a bias error however compared to the BLA. This bias error can be estimated by looking at the noisy BLA and comparing it to the results of the AC analysis. The comparison is shown in Figures 4.4, 4.5 and 4.6. The results at low frequencies are worthless, due to the weak excitation that results from the feedback loop. At higher frequencies, where the signal at the input of the op-amp is larger than the nonlinear noise floor, we can at least say something about the BLA of the op-amp versus the results obtained with the AC analysis. The bias versus variance trade-off mentioned in the paper is clear: The BLA is noisy, but a clear bias error can be seen between the AC FRF and the BLA. Because the op-amp is mainly linear, the error is small, in the order of a few dB for the amplitude and up to a few degrees for the phase.

Besides its use in the spectral correction to determine the nonlinear contribution of a stage, the BLA should also be used to refer the calculated contributions to the output.

$$G_{S,i@Out} = G_{S,i@Stage} \cdot BLA_{Stage\,i \to Out} \tag{4.2}$$

Where $G_{S,i@Out}$ represents the nonlinear contribution of stage i at the output of the op-amp, $G_{S,i@Stage}$ represents the nonlinear contribution of stage i at the output of stage i, calculated using spectral correction. $BLA_{Stage\,i \to Out}$ represents the BLA of the influence of the noise source to the output. The BLA needed here is even harder to determine than the one used in the spectral correction, because there is no excitation present at the frequencies needed to compute the transfer functions to the output. This could be resolved by applying a tiny multisine excitation at the place of the considered noise source during the main transient analysis[4]. This way, the nonlinear operating point is set by the big multisine excitation at the input, while the voltage applied by the noise source(s) allows to calculate the BLA around this nonlinear operating point without altering it.

The multisine in the noise source should excite some even frequency bins and if several small excitations are used, their excited frequency lines should not coincide. If these conditions are satisfied, we can write the contributions to the output at even frequency bins as

$$G_S + BLA_{Stage\,i \to out} \cdot G_N \tag{4.3}$$

where G_S represents the stochastic contribution generated by the nonlinear response to the big multisine excitation at the input, G_N represents the small excitation applied at the location of the assumed (nonlinear) noise source and $BLA_{Stage\,i \to out}$ represents the BLA from stage i to the output. Since the component G_S is stochastic when different phase realisations of the big multisine are considered, averaging over several of these phase realisations can decrease the influence of G_S. In order to estimate its influence, some even bins should be left unexcited.

When this method is applied to the op-amp, both the BLA of the stages and the BLA needed to refer the nonlinear contributions to the output can be determined. This would result in analysis without AC simulations. Applying this method to the op-amp under consideration is a subject for future research.

[4]This technique resembles the one described in [22], where a small excitation was applied on top of a large multisine to determine the out-of-band BLA

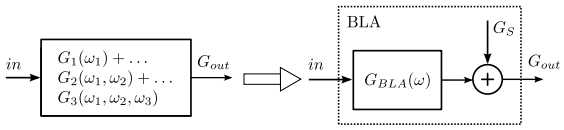


Figure 4.1: Applying the BLA

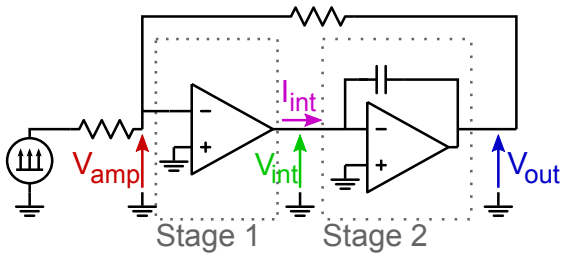


Figure 4.2: Definition of the different voltages and current used in the determination of the BLAs

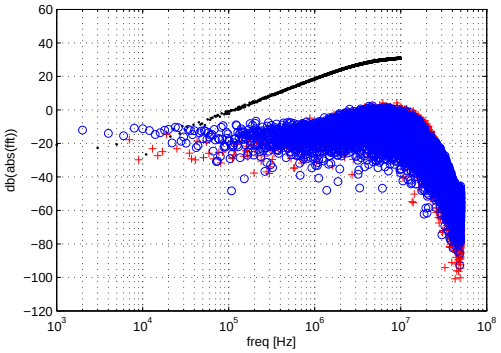


Figure 4.3: Spectrum of the signal V_{amp} when a flat-amplitude random odd random phase multisine excitation is applied to the configuration as shown in Figure 4.2. The symbols $(\cdot)$ represent the excited bins, $(+)$ represent the non-excited odd bins and $(\circ)$ represent the even bins.

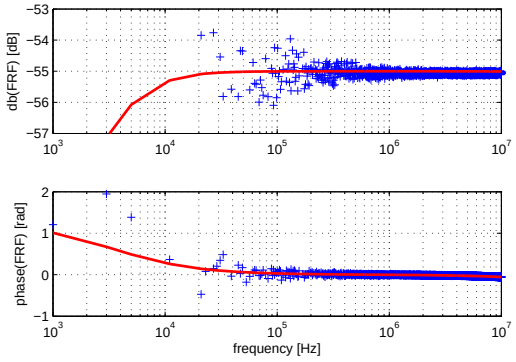


Figure 4.4: Stage 1 voltage to current (I_{int}/V_{amp}) BLA $(+)$ compared to the AC FRF (-)

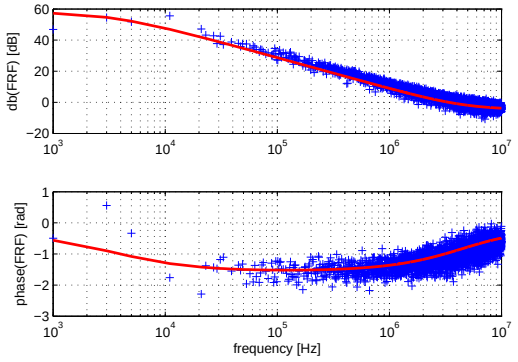


Figure 4.5: Stage 1 voltage to voltage (V_{int}/V_{amp}) BLA $(+)$ compared to the AC FRF (-)

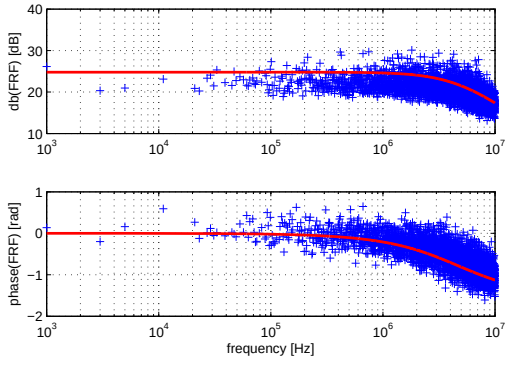


Figure 4.6: Stage 2 voltage to voltage (V_{out}/V_{int}) BLA $(+)$ compared to the AC FRF (-)

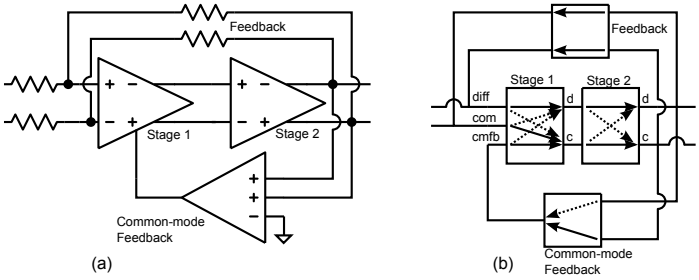


Figure 4.7: (a) The two-stage fully differential op-amp with common-mode feedback in a feedback configuration (b) Schematic representation of the MIMO model of the set-up.

4.2 Nonlinear analysis of a fully-differential op-amp

The op-amp used in the analyses up till now is a single-ended op-amp. To use the nonlinear analysis to improve the performance of fully-differential designs, we need to extend the analysis method.

A schematic representation of the fully-differential op-amp with common-mode feedback is shown in Figure 4.7a. Similar to the paper, we consider a two-stage op-amp, but now including a common-mode feedback circuit.

Stage 1 and 2 now have two inputs and two outputs. We could consider the voltage and current at every input and output, but since we used the differential and common-mode signals during the design of the op-amp, we will also consider them during the analysis.

If we have two nodes with voltages V_1 and V_2, we can always perform the following operation:

$$V_1 = \frac{V_1}{2} + \frac{V_1}{2} + \frac{V_2}{2} - \frac{V_2}{2} = \frac{1}{2}(V_1 + V_2) + \frac{1}{2}(V_1 - V_2) \tag{4.4}$$

$$V_2 = \frac{V_2}{2} + \frac{V_2}{2} + \frac{V_1}{2} - \frac{V_1}{2} = \frac{1}{2}(V_1 + V_2) - \frac{1}{2}(V_1 - V_2) \tag{4.5}$$

It is then possible to define $1/2\,(V_1 + V_2)$ as the common-mode voltage and $1/2(V_1 - V_2)$ as the differential-mode voltage. The same can be done with the currents flowing into node 1 and 2 [9] p. 812. When considering the op-amp from a differential and common-mode point of view, we obtain the model shown in Figure 4.7b. The first stage has three inputs (one for the differential-mode, one for the common-mode and one input for the signal coming from the common-mode feedback stage) and two outputs (differential and common-mode). The second stage consists of two inputs and two outputs (differential and common-mode). The common-mode error amplifier has two inputs (differential and common-mode) and one output.

4.2.1 MIMO representation of the problem

The nodes which are used in the analysis are shown in Figure 4.9. The abbreviations used there will from now on be used in the formulas to denote the node under consideration. All of these nodes (excluding the CMFB node) contain a differential-mode and a common-mode. We will notate the voltages and currents on the node in the following way: First a V for voltage or an I for current, then the name of the node in subscript and finally, DM to denote the differential-mode and CM for the

common-mode. For example, the differential-mode of the output voltage will from now on be called $V_{OUT,DM}$

We will use matrix notation to describe the MIMO transfer functions of the stages. This simplifies the notation of the equations. The frequency response of a MIMO LTI system with n_u inputs and n_y outputs can be written as follows

$$\mathbf{Y}(s) = \mathbf{G}(s) \cdot \mathbf{U}(s) \tag{4.6}$$

where $\mathbf{Y}(s)$ is the $n_y \times 1$ output vector, $\mathbf{U}(s)$ is the $n_u \times 1$ input vector and $\mathbf{G}(s)$ is the $n_y \times n_u$ Frequency Response Matrix, called FRM.

Due to the symmetry of the different stages, an excitation at the common-mode will never have an influence on the differential-mode and vice-versa. Hence the elements which correspond to a signal path going from differential-mode to common-mode and vice-versa in the matrix are zero. These zero entries correspond to the dashed lines from Figure 4.7b.

Asymmetric nonlinear effects like slewing and clipping can introduce a common-mode signal when only a differential excitation is applied. Therefore, under large signal operation, the zero coefficients in the FRM are no longer zero. This is known as a differential to common-mode conversion.

In the paper it was assumed that the system is dominantly linear and the BLA can be approximated by the AC FRF. We confirmed in the previous section that the error made under this assumption remains small. Neglecting the mode conversion (differential-mode to common-mode and vice versa) is a similar assumption. Verifying the assumption is hard though. The full MIMO BLA should be determined in a feedback configuration. This verification is not included in this thesis and is a subject of further research.

The first stage has three inputs and two outputs. resulting in a 2×3 MIMO FRM for each considered frequency. We can write the AC FRM from the input voltages to the output current as

$$\begin{bmatrix} I_{INT,DM}(s) \\ I_{INT,CM}(s) \end{bmatrix} = \begin{bmatrix} S_{1,1}(s) & 0 & 0 \\ 0 & S_{2,2}(s) & S_{2,3}(s) \end{bmatrix} \cdot \begin{bmatrix} V_{AMP,DM}(s) \\ V_{AMP,CM}(s) \\ V_{CMFB}(s) \end{bmatrix} \tag{4.7}$$

$$S_{1,1}(s) = FRF_{V_{AMP,DM} \to I_{INT,DM}}(s) \tag{4.8}$$

$$S_{2,2}(s) = FRF_{V_{AMP,CM} \to I_{INT,CM}}(s) \tag{4.9}$$

$$S_{2,3}(s) = FRF_{V_{CMFB} \to I_{INT,CM}}(s) \tag{4.10}$$

From now on, we will call this 2×3 FRM $\mathbf{S}_1$. The 2×1 vector which contains the differential and common-mode current of the internal node will be called $\mathbf{I}_{INT}$. The reason why the current at the output is considered and not the voltage will be explained later. The 3×1 vector which contains the 3 input voltages of the first stage will be called $\mathbf{V}_{AMP}$. This gives us

$$\mathbf{I}_{INT}(s) = \mathbf{S}_1(s) \cdot \mathbf{V}_{AMP}(s) \tag{4.11}$$

The second stage has two inputs and two outputs. Hence its MIMO FRM consists of a 2×2 matrix. Again, the coefficients of the matrix corresponding to a signal path going from differential to common-mode and vice-versa are zero by the symmetry of the system.

$$\begin{bmatrix} V_{OUT,DM}(s) \\ V_{OUT,CM}(s) \end{bmatrix} = \begin{bmatrix} S_{1,1}(s) & 0 \\ 0 & S_{2,2}(s) \end{bmatrix} \cdot \begin{bmatrix} V_{INT,DM}(s) \\ V_{INT,CM}(s) \end{bmatrix} \tag{4.12}$$

$$S_{1,1}(s) = FRF_{V_{INT,DM} \to V_{OUT,DM}}(s) \tag{4.13}$$

$$S_{2,2}(s) = FRF_{V_{INT,CM} \to V_{OUT,CM}}(s) \tag{4.14}$$

In the same way as before, we define the FRF matrix describing the second stage as $\mathbf{S}_2$, the vector containing the common-mode and differential input voltages at the internal node as $\mathbf{V}_{INT}$ and the vector containing the common-mode and differential output voltages as $\mathbf{V}_{OUT}$. Rewriting the expression in matrix format gives us

$$\mathbf{V}_{OUT}(s) = \mathbf{S}_2(s) \cdot \mathbf{V}_{INT}(s) \tag{4.15}$$

The common-mode feedback stage has two inputs and one output. Because the error-amplifier is built symmetrically with two resistors, the differential-mode will never have influence on the error. Therefore the common-mode feedback stage could be represented with a SISO system, The matrix representation therefore equals

$$V_{CMFB}(s) = \begin{bmatrix} 0 & S^C_{1,2}(s) \end{bmatrix} \cdot \begin{bmatrix} V_{OUT,DM}(s) \\ V_{OUT,CM}(s) \end{bmatrix} \tag{4.16}$$

$$S^C_{1,2}(s) = FRF_{V_{OUT,CM} \to I_{CMFB}}(s) \tag{4.17}$$

we call the FRF matrix describing the common-mode feedback stage $\mathbf{S}_C$ from now on. the vector which contains the output voltages was defined in (4.15) as $\mathbf{V}_{OUT}$. This gives us the following description for the common-mode feedback stage:

$$\mathbf{I}_{CMFB}(s) = \mathbf{S}_C(s) \cdot \mathbf{V}_{OUT}(s) \tag{4.18}$$

Note that the location of the zero coefficients in the matrices $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_C$ allows us to split the total MIMO problem into two SISO problems, one SISO problem for the differential-mode (which is exactly the same as the problem considered in the paper) and one for the common-mode (which contains the two stages and the error amplifier).

4.2.2 Simulation set-up

Because the analysis doesn't work with Ahuja-compensated op-amps, a new two-stage op-amp is designed, The design follows the same design plan as the one described in the previous chapter, but now with Miller-compensated op-amp.

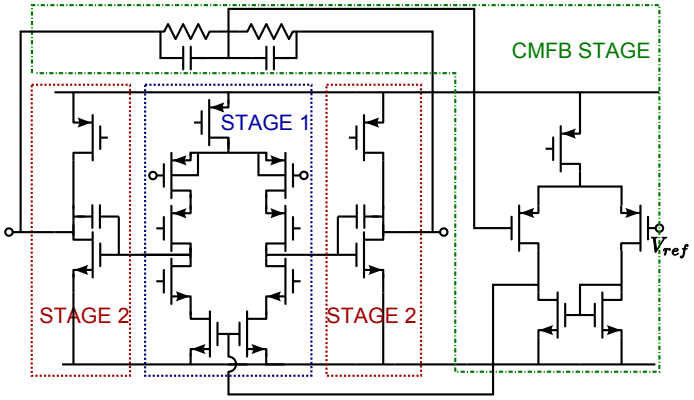


Figure 4.8: Fully-differential Op-amp used in the analysis

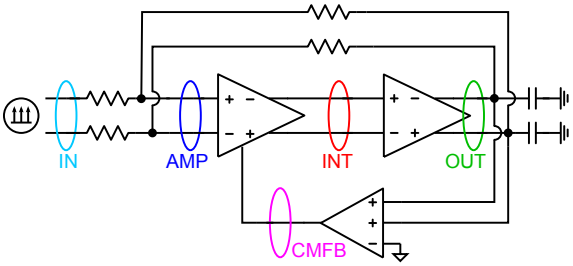


Figure 4.9: Definition of the signals in the two-stage miller-compensated fully differential op-amp with common-mode feedback. All nodes (except CMFB) are transformed into common-mode and differential signals.

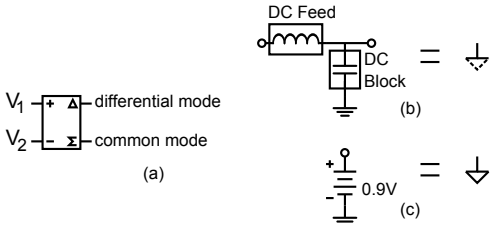


Figure 4.10: Commonly used blocks in the simulation set-ups in ADS. (a) 4 port balun (b) AC ground, as was described in Section 3.8 (c) Signal ground.

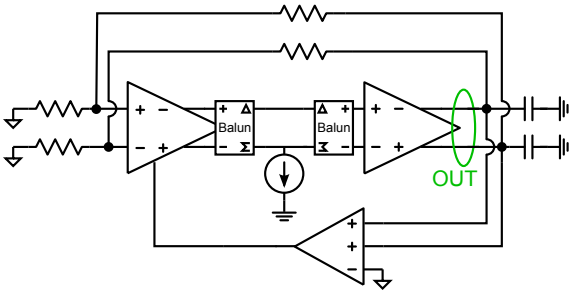


Figure 4.11: Simulation set-up used to determine the influence of the common-mode distortion created by the first stage. The input of the amplifier is placed at AC ground. Using ideal baluns, the signal can be split into common-mode and differential-mode. An AC current source is placed between the common-mode node and ground. By looking at the output during this AC simulation, the noise transfer function is obtained.

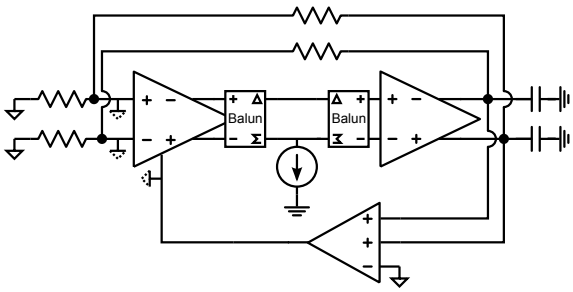


Figure 4.12: Simulation set-up used to determine the total conductance to ground of the common-mode part of the internal node. All inputs of the first stage are AC-grounded. Ideal baluns are used to split the internal signals into a differential and common-mode signal. An AC current source is used to excite the common-mode node. By looking at the response from this current source to the common-mode voltage on the internal node, the total impedance to ground can be obtained.

Instead of a folded-cascode op-amp, a telescopic cascode op-amp is used now. It its architecture and the definitions of the stages are shown in Figure 4.8.
All simulations are performed for this op-amp in an inverting amplifier configuration with gain 10. Two transient analyses are performed. One with a common-mode multisine excitation and another with a differential-mode multisine excitation. From now on they will be referred to as the common-mode experiment and the differential-mode experiment.

Like in the paper, the excitation used is a random odd random phase multisine. It has a frequency resolution of 1kHz and excites odd bins up to 20MHz. The transient analysis time-step corresponds to a sample frequency of 100MHz. The multisine with the smallest crest factor is chosen from 100 realisations and is then scaled in amplitude so that an output voltage swing of 80% of the supply voltage is obtained in the experiment with the differential-mode excitation. The same multisine is used for the common-mode and the differential-mode experiment. Two periods of the multisine are simulated and the first period is discarded to suppress transient effects. Signals of both common-mode and differential-mode are saved and imported in MATLAB for processing.

Performing the transformation to common-mode and differential-mode in MATLAB is not ideal from a numerical point of view. Instead, we will make use of ideal baluns in the simulation set-up to perform this transformation (shown in Figure 4.10a). This way, the simulator can optimise the numerical representation used for the differential and common-mode separately, so that small signals at one of the modes can coexist with large signals on the other mode.

4.2.3 Spectral correction

Now that we know the way we can represent the fully differntial op-amp and we have the results of the transient analyses, we can perform the analysis itself. First, we perform the spectral correction at the output of every stage using equations (4.11), (4.15) and (4.18). This gives us the nonlinear contribution of every stage at its output.

Second stage We will start with the second stage. The vector $\mathbf{G}_{S,2}$ represents the distortion added by the second stage to the common-mode and differential output. The output corrected with the linear response to the input described by equation (4.15) results into

$$\mathbf{G}_{S,2}(s) = \mathbf{V}_{OUT}(s) - \mathbf{V}_{INT}(s) \cdot \mathbf{S}_2(s) \tag{4.19}$$

The noise contributions in the vector can be correlated since we are considering a MIMO problem. We will assume the contributions to be uncorrelated though, because the main source of correlation would be the coupling between the differential and common-mode signals which was assumed to be negligible.

The result of this spectral correction applied to the second stage is shown in Figure 4.13c for the differential-mode experiment and in Figure 4.14c for the common-mode experiment. For the experiment with the differential excitation, the differential-mode of the (nonlinear) noise source contains virtually no component at its even frequency bins due to the symmetry of the circuit. The even frequency bins are therefore not shown in the figure. On the other hand, the common-mode contribution of $\mathbf{G}_{S,2}$ contains no components at the odd bins, so they are also not shown in the figure.

First stage The internal node is loaded by the input conductance of the second stage. In the paper, it was found that if this is the case, it is better to consider the distortion source as a current source between the node to ground. The formula to perform the spectral correction, when a current noise source is assumed, was

$$I_{s,1} = \frac{g_{in,2} + g_{out,1}}{g_{in,2}} \left(I_{meas} - \mathrm{FRF}_{V_{in,1} \to I_{meas}} V_{in,1} \right) \tag{4.20}$$

where $I_{s,1}$ is the nonlinear contribution in the current source, the part between brackets is the spectral correction of the current measured at the output, $g_{in,2}$ is the input conductance of the stage connected to the node and $g_{in,2} + g_{out,1}$ is the total conductance to ground on the considered node.

We can adapt this formula to our situation now. The matrix representation of the part between brackets corresponds to applying spectral correction on the output of stage 1 using (4.11) (that's why the current was considered at the internal node instead of the voltage). The conductances used in (4.20) now depend on the considered mode: the common-mode input impedance of the second stage is different from the differential-mode input impedance of the same stage. The same goes for the total conductance to ground. We can still use AC analyses to determine these conductances though. A schematic representation of the simulation set-up used to determine one of these conductances is shown in Figure 4.12. Performing similar simulations, we obtain the common-mode and differential input conductance of the second stage ($g_{in,diff}$ and $g_{in,comm}$), and the common-mode and differential conductance to ground of the internal node ($g_{tot,diff}$ and $g_{tot,comm}$). To simplify the expressions, we place the fractions in a matrix $\mathbf{G}_{INT}$:

$$\mathbf{G}_{INT} = \begin{bmatrix} g_{tot,diff}/g_{in,diff} \\ g_{tot,comm}/g_{in,comm} \end{bmatrix} \tag{4.21}$$

Using the element-wise product $\langle \circ \rangle$[5], we obtain the MIMO representation of (4.20):

$$\mathbf{G}_{S,1}(s) = \mathbf{G}_{INT}(s) \circ (\mathbf{I}_{INT}(s) - \mathbf{S}_1(s) \cdot \mathbf{V}_{AMP}(s)) \tag{4.22}$$

Using this formula, the nonlinear contributions of the first stage are calculated. They are shown in Figure 4.13a for the differential-mode experiment and in Figure 4.14a for the common-mode experiment. Again, in the differential experiment, the odd bins of the common-mode and the even bins of the differential-mode don't contain any contributions, so they are not shown in the figure.

Common-mode feedback stage The distortion introduced by the common-mode feedback stage remains to be determined. Again, a current noise source has to be considered, because the common-mode feedback circuit is loaded by the input conductance of the third input of the first stage. We can immediately re-use (4.20):

$$\mathbf{G}_C(s) = \frac{g_{tot,CMFB}(s)}{g_{in,Stage1}(s)} (\mathbf{I}_{CMFB}(s) - \mathbf{S}_C(s) \cdot \mathbf{V}_{OUT}(s)) \tag{4.23}$$

[5] $(A \circ B)_{i,j} = (A)_{i,j} \cdot (B)_{i,j}$. This corresponds to using the dotproduct in MATLAB

The input conductance of the common-mode feedback input of the first stage equals two times the gate capacitance of one of the transistors of the first stage. The total input capacitance will therefore be small. At low frequencies, the conductance of this capacitor is so small that numerical precision of the AC analyses came into play. Due to this, when the spectral correction with the impedances was used, the low-frequent contribution became very large and it increases the error at low frequencies. Because of this, the input impedance of the common-mode feedback input of the first stage was assumed infinity and the voltages were used for the spectral correction, instead of the currents. This gives

$$G_C(s) = V_{CMFB}(s) - FRF_{V_{OUT,CM} \to V_{CMFB}}(s) \cdot V_{OUT,CM}(s) \tag{4.24}$$

The result of this spectral correction is shown in Figure 4.13e for the differential-mode experiment and in Figure 4.14e for the common-mode experiment.

4.2.4 Referring to the output

The nonlinear contribution added to the output of every stage are determined above. In order to be able to compare them, we have to refer the contributions to the output. This is done in the same way as before: by determining the FRF from the nonlinear noise source to the output using AC simulations. An AC source is placed at the location of the assumed noise source and its response to the output is calculated. An example of a simulation set-up for a common-mode noise source is shown in Figure 4.11.

Referring the contributions to the output comes down to multiplying the calculated contribution at the stage with the found FRF from source to output. The output referred nonlinear contributions are shown next to the figures which contain the contributions at the output of the stage themselves.

4.2.5 Results: differential-mode experiment

Looking at the results of the experiment with a differential-mode excitation, we see similar results as the ones obtained for the single-ended op-amp: At low frequencies, the nonlinear contributions of the first stage are dominant. At higher frequencies, the second stage starts contributing more tot the total nonlinear distortion, but its contribution never becomes dominant. The trend is similar to the one found with the analysis on a single-ended op-amp. This was to be expected, because the MIMO problem was split into two SISO problems, of which the problem for the differential signal matched the one for the single-ended op-amp.

The first stage is a telescopic cascode stage now, the possible voltage swing will be lower compared to the folded cascode input stage used in the paper. This could explain the reason why now, the contribution of the first stage is dominant over the complete frequency range.

All even order distortion generated by the stages appears in the common-mode signal. When they are referred to the output, the common-mode signals are suppressed by the common-mode feedback. This explains why the distortion at the even frequency bins of $V_{OUT,CM}$ is a lot smaller than the distortion at the differential-mode, while at the stages themselves, about the same amount of even and odd order distortion is generated.

The error made in the analysis can be estimated by looking at the difference between the measured output spectrum and the sum of all the output referred contributions. These comparisons can be seen in Figure 4.15a for the error at the differential-mode of the output signal and in Figure 4.15b for the common-mode of the output signal. Again, the error increases close to the sample frequency due to frequency warping that is present in the trapezoidal method used.

The even bins of the differential-mode and the odd bins of the common-mode are also shown in the figures. It is clear that the signals there are negligible.

4.2.6 Results: Common-mode experiment

For the common-mode experiment, the first stage and the common-mode feedback stage are the main contributors to the distortion. The cancellation of the common-mode signal in the first stage makes that the second stage is only excited with small signals, resulting in low distortion levels. This can also be seen in the comparison between the BLA and the AC FRF of the second stage shown in Figure 4.16b.

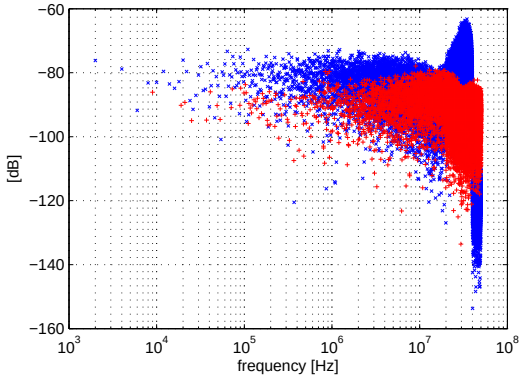
In this experiment, absolutely no contributions were found on the differential-mode of the signals. That's why they are not shown in the figures. This result originates in the fact the simulated system is perfectly symmetrical. If mismatches were to be introduced, results would look differently.

Similar to the differential-mode experiment, we compare the sum of the output referred nonlinear contributions to the measured output spectrum. Figure 4.15d shows the error on the common-mode of the output signal. Again, frequency warping increases the error near the sample frequency. Figure 4.15c shows the differential-mode of the output signal. The distortion level lies around -240dB, and hence is negligible in the differential mode.

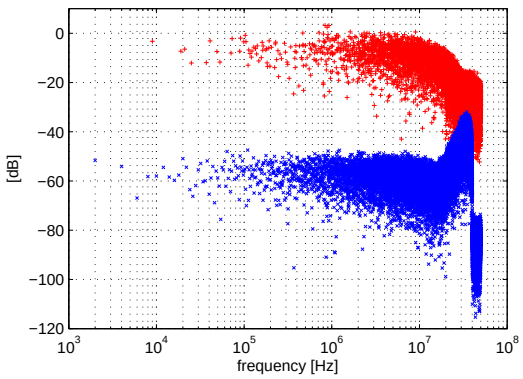
4.3 Improving the nonlinear analysis: Future research

The nonlinear analysis was applied to the fully differential op-amp. This analysis gives a qualitative and quantitative insight in the nonlinear behaviour and design of the circuit. There is however still room for improvement:

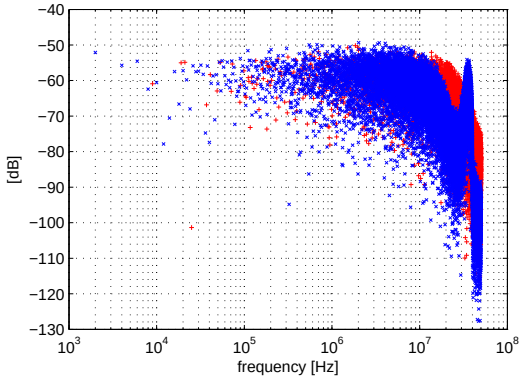
- If the full two-port noise analysis is to be implemented, op-amps with more complicated compensation schemes like the Ahuja-compensation could be analysed.
- Extension to using the full MIMO BLA makes it possible to push the op-amp further into nonlinear operation mode, where slewing and clipping start to have an effect
- Perform a detailed study on the origin of the errors introduced in the analysis method. Correlation between the nonlinear contributions should be investigated. Also a full two-port noise analysis should be implemented to reduce overall errors.
- The number or points on the AC analysis frequency grid can be greatly reduced, lowering the overall simulation time drastically. A correct interpolation method should be applied to keep the interpolation errors reasonably low.



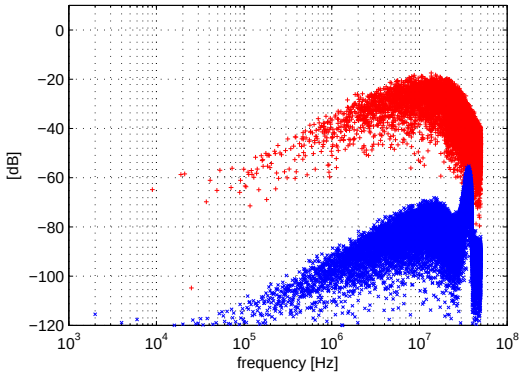
(a) Nonlinear contribution of the first stage at its output.



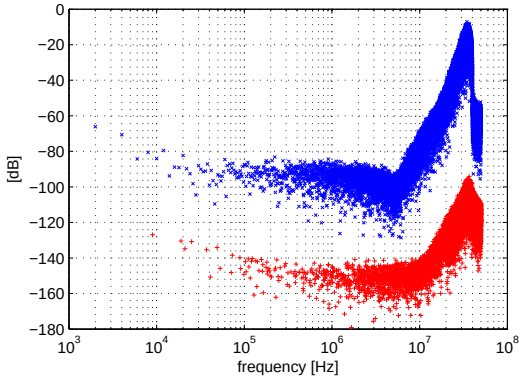
(b) Output referred nonlinear contributions of the first stage.



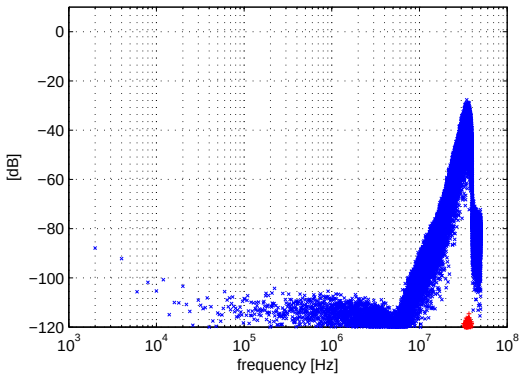
(c) Nonlinear contribution of the second stage at its output.



(d) Output referred nonlinear contributions of the second stage.

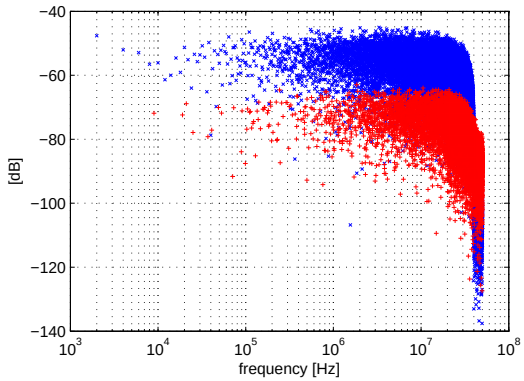


(e) Nonlinear contribution of the common-mode feedback stage at its output. (+) represent the non-excited odd bins (x) are the even bins

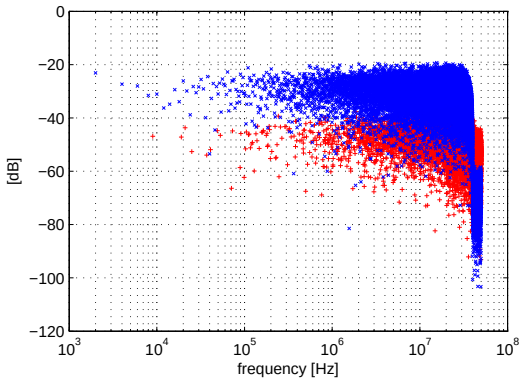


(f) Output referred nonlinear contributions of the common-mode feedback stage. (+) represent the non-excited odd bins (x) are the even bins

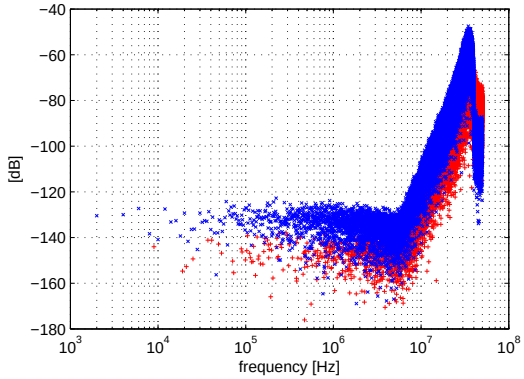
Figure 4.13: differential-mode experiment. In sub-figures a, b, c and d, (+) represent the contributions found on the odd bins of the differential-mode (x) are the contributions found at the even bins of the common-mode. The even bins of the differential-mode and the odd bins of the common-mode of every signal contained no contribution and are therefore not shown.



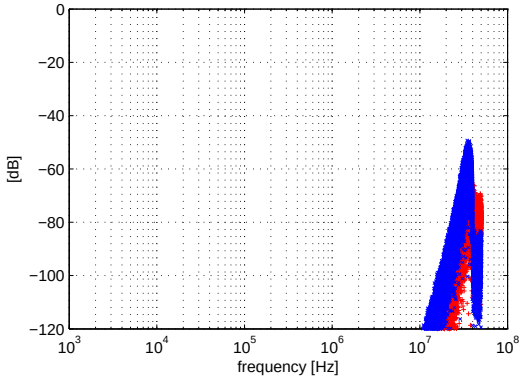
(a) Nonlinear contribution of the first stage at its output.



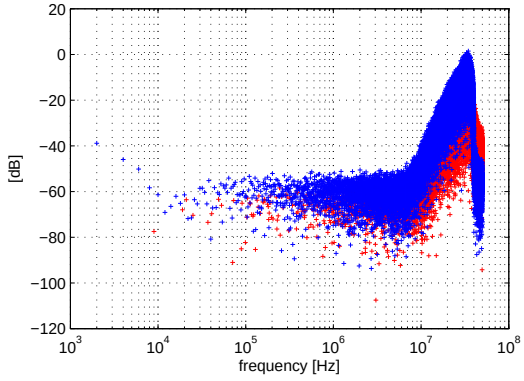
(b) Output referred nonlinear contributions of the first stage.



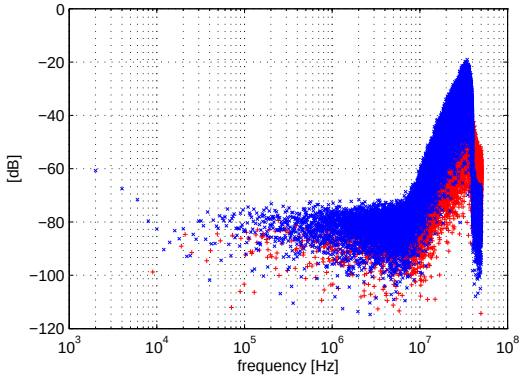
(c) Nonlinear contribution of the second stage at its output.



(d) Output referred nonlinear contributions of the second stage.

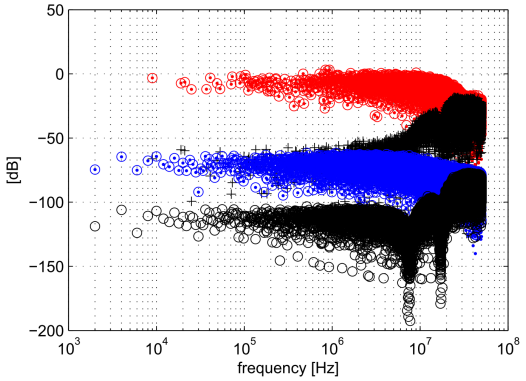


(e) Nonlinear contribution of the common-mode feedback stage at its output.

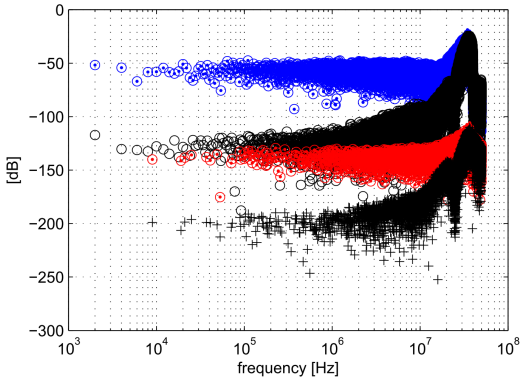


(f) Output referred nonlinear contributions of the common-mode feedback stage.

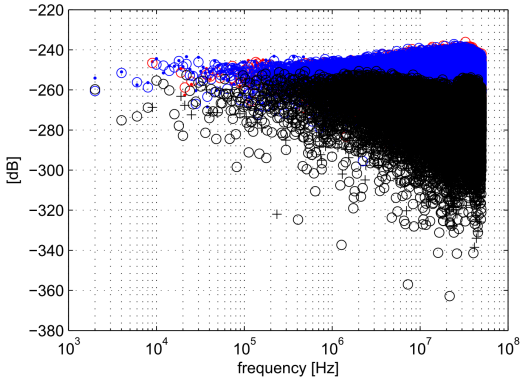
Figure 4.14: Common-mode experiment. (x) are the contributions found at the even bins of the common-mode. (+) represent the contributions found on the odd bins of the common-mode. There were no contributions found at the differential-mode, so they are not shown here.



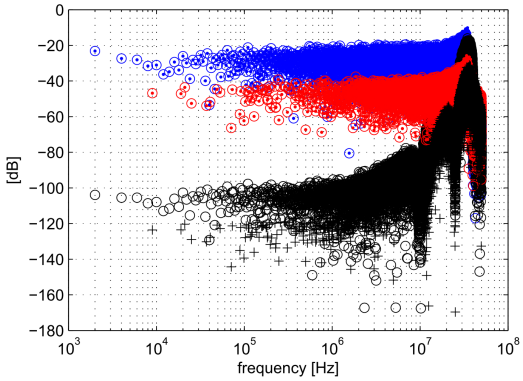
(a) Diffential experiment, differential-mode



(b) Differential experiment, common-mode

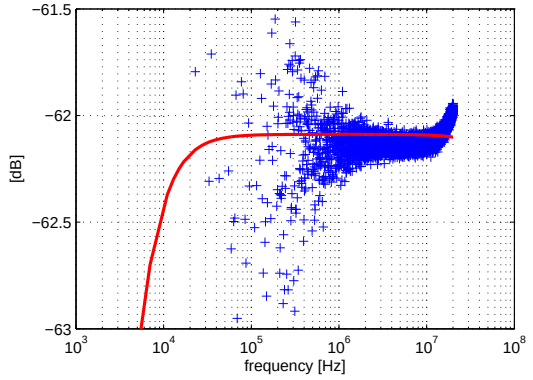


(c) Common-mode experiment, differential-mode

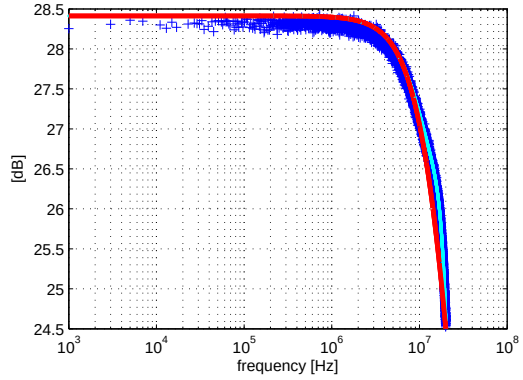


(d) Common-mode experiment, common-mode

Figure 4.15: Comparison between the output spectrum obtained during the transient analysis and the sum of all the output referred nonlinear contributions



(a) First stage, differential-mode of V_{AMP} to differential-mode of V_{INT} during the experiment with an excitation on the differential-mode. (+) BLA (-) AC FRF



(b) Second stage, (+) differential-mode to differential-mode BLA. (+) Common-mode to common-mode BLA and AC FRF. (-) AC FRF

Figure 4.16: Comparison between the BLA and the AC FRF of the stages

Chapter 5

Conclusions

We took some steps in the design of the eSYSID, an electronic SYStem IDentification test-bench. An architecture for the linear dynamic part of the eSYSID was proposed in the first chapter. Robustness of the state-space filter was increased by making the system matrix **A** tridiagonal.

In the second chapter, we chose the MOSFET-C technique to build the (programmable) integrators and the amplifiers needed to represent the state-space equations in an analog way. The transconductor was examined and adapted to produce the least distortion possible. The end of the second chapter describes the calibration circuitry needed to make the filter robust to process and temperature variations.

The third chapter covered the op-amp design. We started with the design of a simple Miller op-amp to demonstrate our design strategy. Then, a more complicated op-amp was chosen to obtain a higher DC gain and the design was made fully differential.

Because nonlinear distortion is the main contributer to the signal to noise and distortion ratio in a MOSFET-C filter, we developed a nonlinear analysis method to pinpoint the dominant nonlinear contributor in the op-amp. First, the method was applied to a single-ended op-amp. Second, the limitations and validity of the hypotheses made during the analysis were verified. Finally the analysis was adapted and applied to a fully differential op-amp.

The combination of nonlinear analysis and the design strategy explained provide us a strong design tool to optimise the nonlinear distortion generated in the complete eSYSID. There is a lot more work to be done before the eSYSID is ready for production though:

Development of the digital programming interface For slowly time-varying systems with large variations in the parameters, the filter must be reprogrammed during the measurement. Because glitches have to be avoided during the measurement, the way the filter is programmed is crucial. Optimisation between programming speed and the number of pins used should be made.

Representation of a wider class of systems In its original specification, the eSYSID contains static nonlinear blocks in order to represent nonlinear systems. For the representation of common electronic circuits, mixers, analog to digital converters (ADC) and digital to analog converters (DAC) are also added. The development of those blocks is left for future research.

Tuning network The architecture of the PLL-based tuning network was discussed at the end of Chapter 2. Although essential for the performance of the global filter, the design of its components was not considered in this thesis.

General optimisation The design of the op-amp was done by hand. Although the hand-designs are sufficient, they could be optimised to improve performance. Using of a good noise analysis, together with our developed nonlinear analysis, the designs of all components could be assessed and improved more efficiently. Mismatches, process variations and yield should also be taken into consideration.

ESD protection If the chip is to be used, protection against Electrostatic Discharge (ESD) is necessary. Its influence on the design should be looked at and minimised.

Lay-out The final step in the design of the chip is the drawing of the lay-out of the components. Because there is a lot of work to be done before the end of the design, the lay-out of the eSYSID was not considered in this thesis.

Improvement of the NL analysis technique The nonlinear analysis elaborated in Chapter 4 could become a powerful design tool if further research is placed into it. Possession of a tool which can be used to determine the dominant source of nonlinear distortion in a complex system would be extremely handy in design. Therefore the analysis method should be applied and adapted for the analysis of other basic analog building blocks, a very interesting source for a lot of future research.

Bibliography

[1] G. Vandersteen, L. D. Locht, and J. Schoukens, "esysid: A proposal for a flexible electronic system identification test bench," in *International Instrumentation and Measurement Technology Conference*, 2009.

[2] D. Turer and R. DeCarlo, "Robust op amp circuit realizations using tridiagonal state space forms," in *Circuits and Systems, 2005. 48th Midwest Symposium on*, aug. 2005, pp. 1 –4 Vol. 1.

[3] T. Kailath, *Linear Systems*. Prentice Hall, 1980.

[4] A. Marthinsen, "On the use of the lanczos method for computing tridiagonal matrices with prescribed eigenvalues," pp. 1–13, 1994.

[5] W. Sansen, *Analog design essentials*. Springer, 2006.

[6] J. Khoury, *Analog Circuits and Devices*. CRC Press, 2003, ch. Continuous-Time Filters, pp. 11.1 – 11.32.

[7] P. Kurahashi, P. K. Hanumolu, G. C. Temes, and U.-K. Moon, "Design of low-voltage highly linear switched-r-mosfet-c filters," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 8, pp. 1699–1709, August 2007.

[8] P. Wambacq, "Analog electronics," 2010-2011, course notes.

[9] P. R. Gray, P. J. Hurst, S. H. Lewis, and R. G. Meyer, *Analysis and Design of Analog Integrated Cricuits*, 4th ed. John Wiley & Sons, Inc, 2001.

[10] Y. Sun, *Design of High Frequency Integrated Analogue Filters*. The Institution of Engineering and Technology, 2002.

[11] Y. Tsividis, "Integrated continuous-time filter design-an overview," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 3, pp. 166–175, March 1994.

[12] Z. Czarnul, "Novel mos resistive circuit for synthesis of fully integrated continuous-time filters," *IEEE Transactions on Circuits and Systems*, vol. CAS-33, no. 7, pp. 718–721, July 1986.

[13] Y. Tsividis, M. Banu, and J. Khoury, "Continuous-time mosfet-c filters in vlsi," *IEEE journal of solid-state circuits*, vol. SC-21, No. 1, pp. 15–30, 1986.

[14] M. Ismael, S. V. Smith, and R. G. Beale, "A new mosfet-c universal filter structure for vlsi," *IEEE journal of solid-state circuits*, vol. 23, no. 1, pp. 183–194, February 1988.

[15] Y. P. Tsividis, *Design of high frequency integrated analogue filters*. The Institution of Engineering and Technology, 2002, ch. The MOSFET-C technique, pp. 35–55.

[16] U.-K. Moon and B.-S. Song, "Low-distortion continuous-time r-mosfet-c filters," in *Circuits and Systems, 1993., ISCAS '93, 1993 IEEE International Symposium on*, may 1993, pp. 1168 –1171 vol.2.

[17] C. Munker, *Phase Noise and Spurious Sidebands in Frequency Synthesizers*, 2005.

[18] Y. Cheng, M. Chan, K. Hui, M. chie Jeng, Z. Liu, J. Huang, K. Chen, J. Chen, R. Tu, P. K. Ko, and C. Hu, *BSIM3v3 Manual*, final version ed., University of California, Berkeley, 1995.

[19] P. Rombouts, "Geavanceed analoog ontwerp," 2011-2012, course notes.

[20] P. Hurst, S. Lewis, J. Keane, F. Aram, and K. Dyer, "Miller compensation using current buffers in fully differential cmos two-stage operational amplifiers," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 2, pp. 275 – 285, feb. 2004.

[21] L. Yao, M. Steyaert, and W. Sansen, "Fast-settling cmos two-stage operational transconductance amplifiers and their systematic design," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 2, 2002, pp. II–839 – II–842 vol.2.

[22] W. Van Moer, Y. Rolain, and A. Geens, "Measurement-based nonlinear modeling of spectral regrowth," *IEEE Transactions on Instrumentation and Measurement*, vol. 50, no. 6, pp. 1711–1716, December 2001.