Chapter 3

Implications of transistor mismatch on analog circuit design and system performance

3.1 Introduction

In a signal processing system several operations or computations are performed on a signal in different stages sequentially. Each of these operations have to emphasize a wanted component or property of the signal without adding too much unwanted extra components. These are due to the non-idealities of the circuit implementation compared to the specified operation. Circuit non-idealities can be divided in two groups: random and systematic errors.

The *random* errors are the result of the stochastic nature of many physical processes. The stochastic behavior of charge carriers in a conductor, for instance, results in various types of noise signals and the stochastic nature of the physical phenomena that take place during the fabrication of integrated circuits, results in a random variation of the properties of the fabricated on-chip devices and mismatches between identically designed devices.

The *systematic* errors occur because a typical circuit implementation only approximates an ideal signal processing operation to a limited extent. These errors are caused, for instance, by the non-linear operating characteristics of devices or by the influence of parasitics in the signal path or device structure.

The effect of these non-idealities can be of different kinds. The *noise* signals limit the minimal signal that can be processed with the system. Device *mismatch* limits the accuracy of the circuit behavior and again limits the minimal signal or energy that is required to execute meaningful signal operation functions. For linear systems, the non-linearities of devices generate *distortion* components of the signals or modulate unwanted 'noise' signals into the used signal band. This typically limits the maximal signal that can be processed correctly.

The circuit designer can reduce the effect of the distortion non-idealities by using small modulation indices for the bias signals; by using large device sizes the impact of mismatch is lowered and by using low impedance levels, the thermal noise signals are reduced. These measures have, however, very important consequences on the power consumption and operation speed of the system. Therefore the quality of a circuit realization is evaluated from the obtained accuracy, noise level or linearity relative to the used power and the speed of operation. The designer will try to achieve for a given speed the best performance with a minimal power consumption.

The fundamental impact of noise on the overall system performance has been studied extensively in literature see e.g. [Vit 90b] [Voo 93, and its references]. In this chapter we investigate the impact of transistor mismatch on the total performance of analog circuits and systems. First, we discuss the characterization and modeling of transistor mismatch and describe a new extraction method to derive the matching quality of sub-micron CMOS technologies. This quantitative model information is very important for the design of analog circuits since it allows the designer to accurately predict the accuracy performance. Furthermore, it forms the basis for the evaluation of the impact of transistor mismatch on the analog performance.

The implications of transistor mismatch on the design of basic analog building blocks is then discussed in detail in sections 3.3 and 3.4. The speed, accuracy and power consumption performances of analog circuits are linked due to the effect of mismatch on the circuit design; guidelines for the optimal design of circuits are derived. In section 3.5 we generalize these results and prove that mismatch puts a fundamental limitation on the maximal total performance of analog signal processing systems. The $Speed{\cdot}Accuracy^2/Power$ ratio is fixed by technological constants that express the matching quality of the technology. For circuit building blocks with high accuracy requirements thermal noise is considered as the limiting factor for performance improvement or power consumption reduction [Vit 94, Dij 94] but we show that the impact of transistor mismatch on the minimal power consumption is more important for present-day CMOS technologies than the impact of thermal noise for high speed analog circuits and massively parallel analog systems. The matching performance is technology dependent and the scaling of the circuit performance with the down-scaling of the technology size is discussed in section 3.5.4. The techniques to reduce the impact of mismatch and their effect on the performance of systems is reviewed in section 3.6. The last section (3.8) treats the implications for the VLSI design of massively parallel analog systems and discusses the advantages of analog VLSI implementation over the digital VLSI implementation for massively parallel analog systems.

3.2 Modeling and characterization of transistor mismatch

3.2.1 What is transistor mismatch

Two identical designed devices on an integrated circuit have random differences in their behavior and show a certain level of random mismatch in the parameters which model their behavior. This mismatch is due to the stochastic nature of physical processes that are used to fabricate the device. In [Pel 89] the following definition for mismatch is given: *mismatch is the process that causes time-independent random variations in physical quantities of identically designed devices.*

3.2.2 Modeling of CMOS transistor mismatch

Mismatch in device parameters can be modeled by using different techniques. Several authors [Laks86] [Shy 84] [Miz 94] start from the physical background of the parameters to calculate and model the device mismatch dependence on technology parameters and device size. Or a black box approach can be used by supposing the statistical properties of the different mismatch generation processes and calculating their influence on different circuit parameters [Pel 89].

The mismatch of two CMOS identical transistors is characterized by the random variation of the difference in their threshold voltage V_{T0}, their body factor γ and their current factor β (the definitions of these parameters can be found in appendix A). For technologies with a minimal device size larger than typically 2 μm, a widely accepted and experimentally verified model [Pel 89] [Bas 95] [Pav 94] for these random variations is a normal distribution with mean equal to zero and a variance dependent on the gate-width W and gate-length L and the mutual distance D between the devices:

$$\sigma^2(\Delta V_{T0}) = \frac{A_{VT0}^2}{WL} + S_{V_{T0}}^2 D^2 \tag{3.1}$$

$$\sigma^2(\Delta\gamma) = \frac{A_\gamma^2}{WL} + S_\gamma^2 D^2 \tag{3.2}$$

$$\left(\frac{\sigma(\Delta\beta)}{\beta}\right)^2 = \frac{A_\beta^2}{WL} + S_\beta^2 D^2 \tag{3.3}$$

A_{VT0}, A_γ, A_β, $S_{V_{T0}}$, S_γ and S_β are process-dependent constants. In table 3.1 and 3.2 the proportionality constants for several processes are summarized. Experimental data show that the correlation between the V_{T0} and β mismatch is very low although both parameters depend on the oxide thickness[Pel 89] [Bas 95].

The last two columns of table 3.2 contain corner distances at which the distance dependent term in the parameter mismatch becomes dominant over the size dependent term. The corner distances D_m is defined as the distance for which the mismatch due to the distance effect on a parameter p is equal to the mismatch due to the size dependence for a minimal size device ($W = \lambda_T$ and $L = \lambda_T$, where λ_T is the minimal size of the technology) and is calculated as:

$$D_m = \frac{A}{\lambda_T \cdot S} \tag{3.4}$$

For devices with an area of A_f times the minimal area the critical distance D_m is $\sqrt{A_f}$ times smaller. The obtained critical distances D_m for the present-day processes are very

Technology	Type	A_{VT0}	A_β	$(V_{GS} - V_T)_m$
		[mVμm]	[%μm]	[V]
2.5μm [Pel 89]	nMOS	30	2.3	2.6
	pMOS	35	3.2	2.2
1.2μm [Bas 95]	nMOS	21	1.8	2.3
	pMOS	25	4.2	1.2
0.7μm	nMOS	13	1.9	1.4
	pMOS	22	2.8	1.6

Table 3.1: The matching proportionality constants for size dependence for different industrial CMOS processes. The parameter $(V_{GS} - V_T)_m$ is defined in (3.33) in section 3.3.

Technology	Type	$S_{V_{T0}}$	S_β	D_{VTm}	$D_{\beta m}$
λ_T		[μV /μm]	[ppm/μm]	[mm]	[mm]
2.5μm [Pel 89]	nMOS	4	2	3	5
	pMOS	4	2	3.5	13
1.2μm [Bas 95]	nMOS	0.3	3	58	5
	pMOS	0.6	5	35	12
0.7μm	nMOS	0.4	2	46	14
	pMOS	-	3	-	13

Table 3.2: The matching proportionality constants for distance dependence for different industrial CMOS processes.

large compared to the typical size of an analog circuit. Therefore the distance dependence of the parameter mismatch will be neglected in the discussion of the impact of transistor mismatch on analog circuit and system performance.

In equations (3.1), (3.2) and (3.3) the standard deviation of the difference of the parameters of two transistors is given. For a random variable Z defined as $Z = X - Y$ the variance is $\sigma^2(Z) = \sigma^2(X) + \sigma^2(Y)$. The following relations are thus obtained for the variance of the absolute parameters of a single transistor:

$$\sigma(V_{T0}) = \sqrt{2}\sigma(\Delta V_{T0}) \tag{3.5}$$

$$\sigma(\gamma) = \sqrt{2}\sigma(\Delta\gamma) \tag{3.6}$$

$$\left(\frac{\sigma(\beta)}{\beta}\right) = \sqrt{2}\left(\frac{\sigma(\Delta\beta)}{\beta}\right) \tag{3.7}$$

3.2.3 Characterization of transistor mismatch

The matching behavior of transistors is very strongly dependent on the used IC technology. Therefore an in-house characterization procedure has been set-up [Bas 95]. A new direct extraction algorithm has been developed to extract the ΔV_{T0} and $\left(\frac{\Delta\beta}{\beta}\right)$ of a transistor pair from their measured relative current difference $\left(\frac{\Delta I_{DS}}{I_{DS}}\right)$ in saturation. The big advantage of measuring currents in saturation is the much lower sensitivity to parasitics in the set-up, which becomes more and more important for sub-micron and deep sub-micron technologies. Also, the model of V_{T0} mismatch of minimal sized devices in sub-micron technologies has been improved. In this paragraph a short overview is given; for more details the reader is referred to [Stey94b] [Bas 95], [Bas 96c], [Bas 96b], and [Bas 96a].

A) Test Circuits Test-circuits are processed to experimentally check the validity of the models and to determine the proportionality constants of the size and distance dependence of transistor mismatch. In figure 3.1 a micro-photograph of the nMOS test-chip for a 1.2 μm CMOS technology is presented. The test-chip contains a matrix of transistors. On a row identical transistors are spaced at different mutual distances to examine the mismatch spatial dependence. The different rows contain transistors with different sizes to determine the mismatch dependence on device size. All sources are connected to a common point. Transistors in the same row have their gates connected and transistors in the same column have their drains connected. Special attention has to be paid in the layout to obtain a very low resistance in the source path to eliminate systematic errors during the measurements; very wide source metal connections are used and can be clearly distinguished in figure 3.1. Two separate test-chips for the characterization of the nMOS transistors and pMOS transistors have been designed.

B) Measurement set-up The measurements are carried out on packaged test-circuits using a HP4062A Semiconductor Parametric Test System including a switch matrix and

Figure 3.1: Micro-photograph of the mismatch characterization test-chip in a 1.2 μm CMOS technology.

two voltage source units. The drain current of the different transistors in a row are accessed sequentially through a switch matrix which connects the drain of the transistor under test to the current meter; the drains of the other transistor in the row are left open. The switch-matrix connects the gate of the transistor under test to the gate voltage source and connects the gates of the other rows to ground. The transistors are biased in strong inversion by using gate voltages larger than V_T; the current is measured in saturation by applying a constant drain voltage larger than the maximal $(V_{GS} - V_T)$. The drain voltage is applied by using a *4 point technique* where two separate *sense* wires, which carry no current, monitor the drain voltage and the current flow is through two separate *force* wires. For the C12 technology, for instance, the gate voltage is swept in 26 steps from 0.75 V to 2.0 V with a constant drain voltage of 2.0 V. Since the different transistors are measured sequentially the DC repeatability of the DC gate voltage source must be larger than the smallest gate-voltage mismatch we want to measure. The repeatability of the source in our set-up was better than 6 digits which is more than sufficient.

For the extraction of the transistor mismatch we are interested in the current differences between the different transistor pairs. The current difference can only be obtained by measuring the currents separately and then subtracting the current measurements. This procedure is very sensitive to errors, but it is the only way to obtain a measurement of the current difference of two transistors. To have an accurate estimation of the current difference, the individual currents have to be measured very accurately. The necessary relative accuracy on the the drain current measurement $\left(\frac{\sigma_m(I_{DS})}{I_{DS}}\right)$ is dependent on the relative current difference $\left(\frac{\Delta I}{I}\right)$ we want to measure and on the wanted relative accuracy for the current difference $\left(\frac{\sigma_m(\Delta I_{DS})}{\Delta I_{DS}}\right)$:

$$\left(\frac{\sigma_m(I_{DS})}{I_{DS}}\right) = \frac{1}{\sqrt{2}}\left(\frac{\Delta I}{I}\right)\left(\frac{\sigma_m(\Delta I_{DS})}{\Delta I_{DS}}\right) \tag{3.8}$$

The necessary number of digits in the current measurement must be larger as the number of digits we want in the current difference measurement plus the relative current difference we are measuring expressed in a number of digits. In our set-up a HP3457 multi-meter with a 5 to 7 1/2 digits resolution is used.

Using this procedure the current through all transistors in the array as function of the gate voltage is measured and stored.

C) New Mismatch Parameter Extraction Technique The relative current difference of a transistor pair is dependent on both the current factor β matching, the threshold voltage V_{T0} matching and the gate voltage; in section 3.3.2 the following dependence is derived for transistors biased in saturation:

$$\left(\frac{\Delta I_{DS}}{I_{DS}}\right) = \left(\frac{\Delta\beta}{\beta}\right) - \frac{2\Delta V_{T0}}{(V_{GS} - V_T)} \tag{3.9}$$

Since the current measurements are performed in saturation, the current of a transistor is in first order not dependent on the drain voltage but only on the gate-source voltage.

Parasitic resistances in the drain path, from e.g. the switches in the switch-matrix or the current mirror, do not generate errors. However, differences in the parasitic resistors in the source path of the two devices under test generate systematic errors; therefore wide interconnects are used on the test chip in the source connections of the transistors.

First the threshold voltage is extracted for one of the transistors using a standard extraction technique [Bas 95]. Then the model of (3.9) is fitted to the relative current difference measurement using a linear least-squares algorithm. The mismatch parameters ΔV_{T0} and $\left(\frac{\Delta\beta}{\beta}\right)$ are obtained directly from the measured current difference.

Other techniques can be used to calculate the V_{T0} matching and β matching. In [Pel 89] the transistors are measured in the linear region and for each individual transistor the threshold voltage V_{T0} and the current factor β are extracted using classical parameter extraction algorithms. The parameter mismatch is then calculated by subtracting the parameters of the individual transistors.

Also a direct extraction technique can be derived from a more complex drain current model which includes a mobility reduction parameter θ. The extra mismatch parameter $\Delta\theta$ is however highly correlated to the $\left(\frac{\Delta\beta}{\beta}\right)$ and the overall modeling of the current mismatch becomes more complicated without improvements in accuracy [Bas 95].

We conclude that he new direct extraction procedure has two main advantages:

- The current measurements are performed in *saturation* so that parasitic resistances in the drain path do not generate errors. In sub-micron technologies where the current factor β is large due to the thinner oxide, the equivalent resistance of the transistors biased in the linear region becomes very small so that any parasitic series resistance in the current path gives rise to important errors. Thus especially for sub-micron technologies measuring in saturation region is a more robust technique. Moreover, the majority of the transistors are biased in saturation in analog design so that the mismatch parameters are obtained under realistic conditions.

- The parameter mismatches are extracted *directly* from the current difference measurements with a high accuracy; when first the absolute parameters are extracted and the mismatch parameters are calculated as a difference of absolute parameters, a very high accuracy in the absolute parameters is necessary to obtain accurate mismatch parameters - the same calculations (3.8) as for the current measurements can be used to calculate the necessary accuracy of the absolute parameters. A higher accuracy can be obtained in the individual absolute current measurements by using highly accurate measurement equipment than in the extraction of the absolute parameters and thus the presented direct technique gives more accurate results [Bas 95].

In this way a measurement of the ΔV_{T0} and $\left(\frac{\Delta\beta}{\beta}\right)$ is obtained for each of the transistor pairs, which all have different mutual distances and different gate areas, for every test-chip. From this experimental data the proportionality constants for the mismatch dependence on distance and on size can be extracted for the models in (3.1) and (3.3).

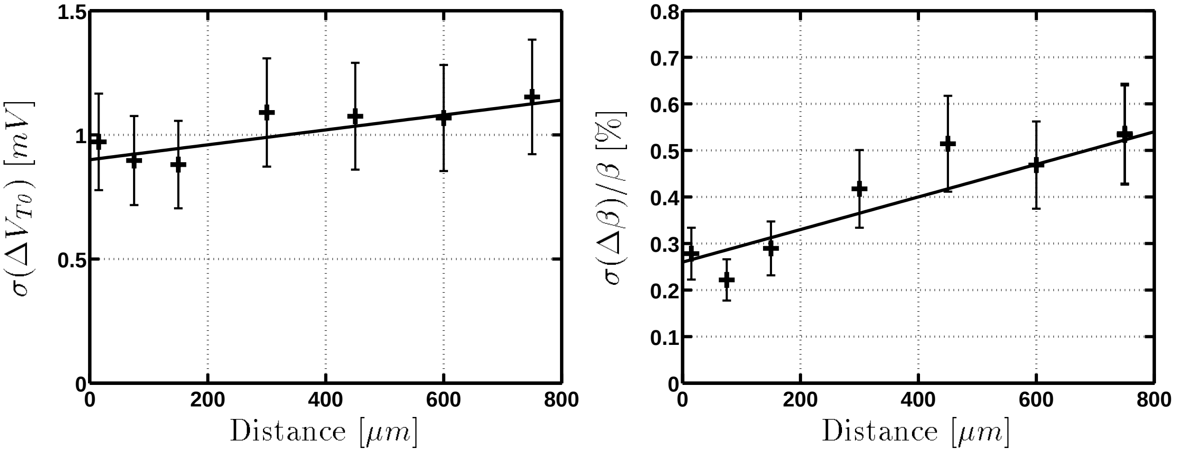
D) Mismatch Dependence on Distance For every row - containing transistors of the same size - the first transistor is used as a reference and the ΔV_{T0} and $\left(\frac{\Delta\beta}{\beta}\right)$ are extracted for the consecutive pairs as a function of the distance. For every size and every distance one sample is obtained per test-chip. Then the standard deviation of the mismatch parameters $\sigma(\Delta V_{T0})$ and $\left(\frac{\sigma(\Delta\beta)}{\beta}\right)$ is calculated by combining the samples of all test-chips and calculating the sample variance from the MAD (median of the absolute differences) [Rey 83] to eliminate the effect of outliers.

At this point it is important to discuss the accuracy of the extracted standard deviations. The estimation of the standard deviation is an application of the estimation of a parameter of the distribution of a random variable [Pap 91]. For a normally distributed random variable, the sample variance s^2 can be used as an estimation for the variance σ^2 and the s^2/σ^2 ratio follows a Chi-squared distribution. This allows to determine the confidence interval for the extracted value for a given confidence level (see also [Per 95]). In our extractions we have aimed at a ±20% accuracy with a 99.7% confidence, which requires a sample size of over 100 samples. For the same confidence level, over 500 samples or test-chips would be required to attain an accuracy of ±10%. These numbers clearly illustrate, the very high measurement effort that has to be done to obtain good quantitative mismatch parameters.

For small transistors the distance effect is completely masked by the large variance due to the small gate area as can be noted in (3.1) and (3.3) and from table 3.2. In the 1.2 μm CMOS technology, for instance, a significant distance dependence is only observed for a 20/20 $\mu m/\mu m$ nMOS transistor as is shown in figure 3.2. A straight line is fitted to the standard deviation data points and the $S_{V_{T0}}$ and S_β from (3.1) and (3.3) are extracted. In table 3.2 the distance dependence model parameters for several technologies are summarized.

E) Mismatch Dependence on Size The rows contain transistors of different sizes so that the size dependence of the parameter mismatch can be investigated. For each transistor size, the standard deviation of the parameter mismatch is again estimated from the sample obtained by combining the results of the devices at minimum distance over all test-chips. The same statistical techniques are used as for the distance dependence.

Threshold voltage V_{T0} mismatch In figure 3.3 the standard deviation of the threshold voltage mismatch is plotted versus the square root of the effective area for the 10 transistor sizes on the C12 test-chip. The model of (3.1) predicts a linear relation between the $\sigma(V_{T0})$ and the square root of the effective area. For the large transistors, which do not have a minimal width nor a minimal length, the experimental results confirm the model; this also agrees with the experimental results from other authors [Pel 89, Laks86, Mic 92]. A straight line is fitted and the size proportionality constant is obtained (see table 3.1).



(a) (b)

Figure 3.2: Threshold voltage mismatch (a) and current factor mismatch (b) for the 20/20 nMOS transistor versus distance in the 1.2 μm CMOS technology; a straight line is fitted through the measurement points to extract the mismatch distance dependence.

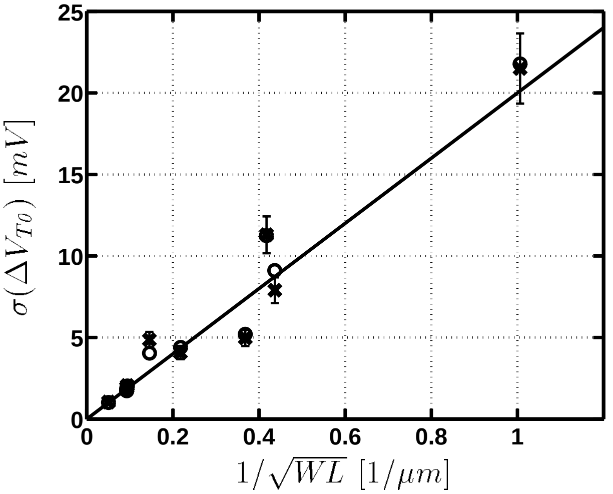


Figure 3.3: The standard deviation of the threshold voltage V_{T0} mismatch versus the square root of the gate area for nMOS transistors in the 1.2 μm CMOS technology; the experimental results are represented by + and the values predicted by the model in equation (3.10) are represented by o; the straight line represents the predictions by the linear model of equation (3.1).

Accurate modeling of V_{T0} mismatch in sub-micron technologies Narrow channel transistors with a minimal width ($W = 1.4\mu m$, $L = 6.2\mu m$ and $1/\sqrt{W_{eff} L_{eff}} = 0.37/\mu m$) show less mismatch than predicted by the linear model (3.1) as can be verified from the experimental results in figure 3.3; short channel transistors with a minimal gate length ($L = 1.2\mu m$, $W = 6.2\mu m$ and $1/\sqrt{W_{eff} L_{eff}} = 0.42/\mu m$; or $L = 1.2\mu m$, $W = 50\mu m$ and $1/\sqrt{W_{eff} L_{eff}} = 0.15/\mu m$) on the other hand show a significant higher mismatch than predicted by the linear model. In high speed analog designs, the designer prefers to use small gate-lengths so that the highest intrinsic speed f_T for the transistor is obtained [Lak 94]; accurate models for minimum sized transistors are thus necessary.

For the accurate modeling of the threshold mismatch in sub-micron technologies the simple linear model has to be extended for short and narrow channel effects. The threshold voltage is dependent on the flat-band voltage, the surface potential, the depletion charge and the gate capacitance [Lak 94]. It has been verified experimentally that the mismatch of the threshold voltage is mainly attributed on the mismatch of the bulk depletion charges in the two devices [Pel 89, Miz 94, Laks86]. Due to the random process of the ion implantation and the drive-in diffusion process, the doping ions are distributed randomly and the depletion charge fluctuates randomly. The depletion charge follows a Poisson distribution: the mean of the depletion charge is proportional to the gate-area and the bulk doping level; the variance is equal to the square root of the mean of the depletion charge. This leads to the linear model (3.1) where the standard deviation of the threshold voltage is inversely proportional to the gate area.

In sub-micron technologies two effects introduce errors in the model. Due to the presence of the source and the drain diffusion areas and the charge sharing effect, part of the channel depletion charge is not controlled by the gate voltage anymore. For devices with a small gate-length, this charge is a relatively large part of the depletion charge. The depletion charge controlled by the gate is smaller and as a result, the threshold voltage lowers for small gate lengths whereas the variance of the threshold voltage or the V_T mismatch increases.

The depletion charge is not limited to the gate area but due to the fringing field some of the dopant atoms on the side are also depleted. For large widths, the part of the depletion region on the sides is a small percentage of the total depletion region volume. But for narrow channel devices, the side parts are a large percentage of the depletion charge. The depletion charge controlled by the gate is now larger so that the threshold voltage increases and the V_T mismatch decreases for narrow channel devices.

The narrow and short channel effects explain the deviations of the V_T mismatch from the linear model in the experimental data very well. When these effects are modeled quantitatively, the following extended model for the V_{T0} mismatch is obtained [Stey94b] [Bas 95]:

$$\sigma^2(\Delta V_{T0}) = \frac{A_{1VT}}{WL} + \frac{A_{2VT}}{WL^2} - \frac{A_{3VT}}{W^2L} \qquad (3.10)$$

In this extended model the second term models the short channel effect for transistors with small gate-lengths; the last term results in a lower mismatch for small gate-widths

parameter	nMOS	pMOS	
A_{1VT}	20	23	$mV\ \mu m$
A_{2VT}	19	20	$mV\ \mu m^{3/2}$
A_{3VT}	18	12	$mV\ \mu m^{3/2}$

Table 3.3: Mismatch fitting constants for the extended model of equation (3.10) for a 1.2 μm CMOS technology.

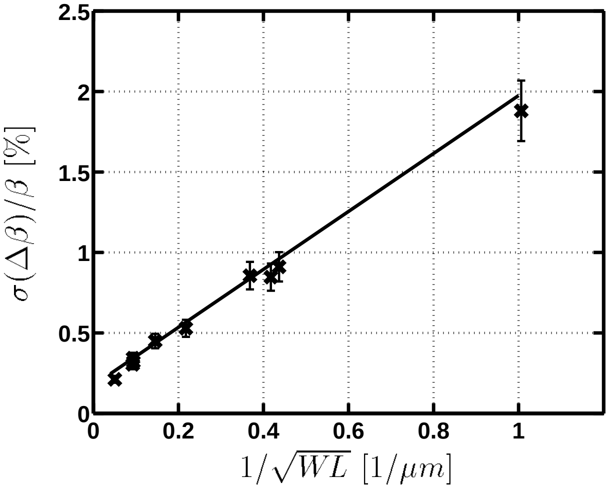


Figure 3.4: The standard deviation of the current factor mismatch as function of the square root of the gate-ares for nMOS devices in the 1.2 μm CMOS technology.

and models the narrow channel effect. This new model is able to predict the mismatch data within the confidence limits. The model parameters for the 1.2 μm technology are summarized in table 3.3 [Bas 95] and in figure 3.3 the results for the fitting of the new model are represented by the circle; a very good agreement with the experimental data is obtained.

Current factor β mismatch In figure 3.4 the standard deviation of the current factor mismatch is plotted as function of the square root of the effective area for nMOS devices in the C12 technology. The experimental data fits well to the linear model of (3.3) and no significant deviation for short or narrow channel devices is observed.

F) Extraction Validation The correlation factor between the V_T and β mismatch also has to be computed. In all characterized technologies the correlation factor remains very low and the correlation can be neglected [Bas 95]. This agrees well with the experimental results of other authors [Pel 89, Laks86]. To check the accuracy of the characterization procedure, the measured and the predicted current mismatch are compared. In saturation the predicted variance of the current mismatch is given by (3.29) and (3.31), when the correlation between the parameters is negligible. The standard deviation of the measured

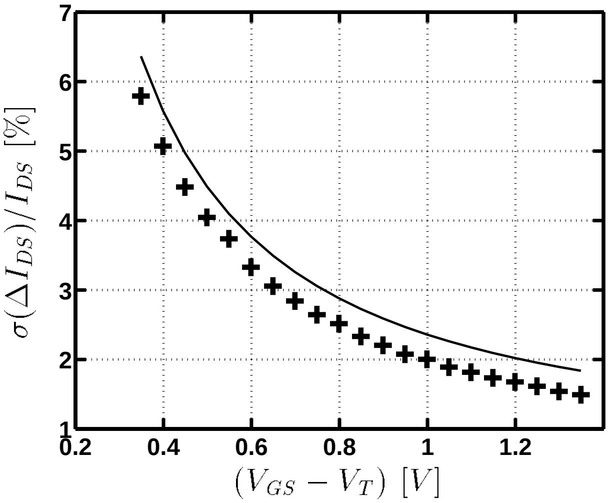


Figure 3.5: Standard deviation of the relative current mismatch versus the gate-overdrive voltage for a 6.2/1.2 nMOS transistor in a 1.2 μm CMOS technology; the crosses indicate the measurements and the solid lines are the theoretical predictions from the extracted mismatch in the parameters.

current mismatch, for each bias point and for each device size, has been calculated and is compared with the predicted value. In figure 3.5 the measured and predicted standard deviation of the current mismatch is plotted for the 6.2/1.2 $\mu m/\mu m$ nMOS transistor in the 1.2 μm technology is plotted. The current mismatch is predicted within 20% over the complete measurement range which is within the accuracy limits of the experimental data.

Summary We conclude that the characterization of transistor mismatch is a tedious process which requires a very large measurement effort. The design, realization and validation of the measurement set-up, the acquisition of the experimental data and the statistical processing of the data have to be performed with great care to avoid errors and systematic effects. Moreover, when migrating towards sub-micron and deep-sub-micron technologies, the standard mismatch models have to checked for their validity; if necessary the effects of the short or narrow channel effects have to be accounted for in model extensions.

On the other hand, in the rest of this chapter, we will demonstrate the importance of a good knowledge of the matching behavior of devices. A circuit designer can only improve the accuracy of circuits by increasing the device area of the devices. Unfortunately, the capacitive load in circuits is also proportional to the area. Generating a signal excursion across a capacitor, results in loading and unloading currents so that the circuit dissipates energy proportional to the capacitive load or the accuracy. The quality of the device matching thus is the ultimate limit for the power consumption of circuits; the impact of transistor mismatch is even more important in high speed applications than thermal noise.

The better the mismatch of devices is modeled and characterized, the smaller area's the designer can safely use while keeping a high circuit yield; consequently the circuits will consume less power for the specified accuracy and speed. When no mismatch data is available, very conservative designs using large devices have to be used and very poor overall circuit performance is achieved.

3.3 Implications of mismatch on transistor behavior

In this section the consequence of parameter mismatch on the transistor behavior is calculated for the different possible operation regions of the transistor. First, we introduce the necessary mathematical techniques for mismatch calculations.

3.3.1 Mathematical techniques

Transistor mismatch results in small deviations of the transistor parameters from their nominal value and these deviations result in small deviations of the circuit characteristics from their nominal values. In order to calculate the effect of transistor mismatch on circuit characteristics the following relations are used.

For a circuit characteristic Z, which is defined as $Z = f(x, y)$, the deviation δ_Z in Z due to a deviation δ_x in x and δ_y in y is calculated as:

$$\delta_Z = \frac{\partial f}{\partial x}\delta_x + \frac{\partial f}{\partial y}\delta_y \tag{3.11}$$

For independent and normally distributed deviations in x and y the standard deviation of Z is:

$$\sigma^2(Z) = \left(\frac{\partial f}{\partial x}\right)^2 \sigma^2(x) + \left(\frac{\partial f}{\partial y}\right)^2 \sigma^2(y) \tag{3.12}$$

For two characteristics Z_1 and Z_2 defined as $Z_1 = f(x_1, y_1)$ and $Z_2 = f(x_2, y_2)$ the following relations are derived applying (3.11) and (3.12), with $\Delta x = x_1 - x_2$ and $\Delta y = y_1 - y_2$:

$$\Delta Z = Z_1 - Z_2 = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y \tag{3.13}$$

$$\sigma^2(\Delta Z) = \left(\frac{\partial f}{\partial x}\right)^2 \sigma^2(\Delta x) + \left(\frac{\partial f}{\partial y}\right)^2 \sigma^2(\Delta y) \tag{3.14}$$

$$\sigma(\Delta Z) = \sqrt{2}\sigma(Z) \tag{3.15}$$

These relations allow us to calculate the effect of parameter mismatches and transistor mismatches on the transistor or circuit behavior.

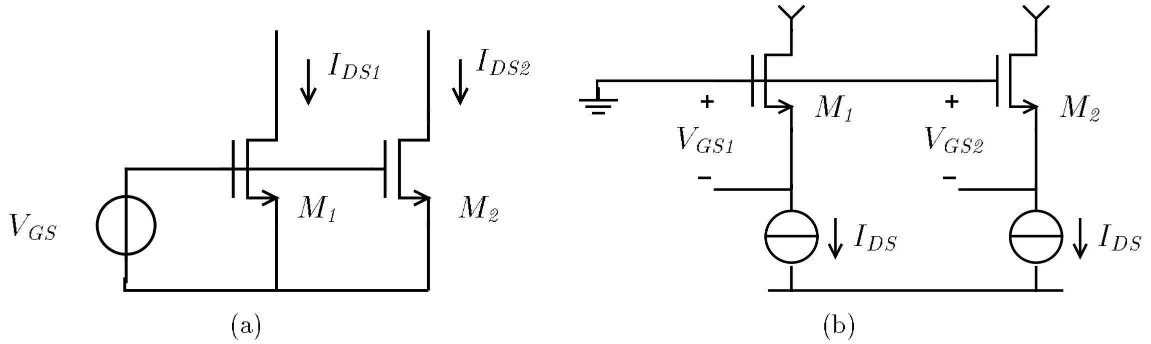


Figure 3.6: (a) Two transistors biased with equal gate-source voltage; (b) two transistors biased with equal drain-source current.

3.3.2 Implications on transistor behavior

A circuit designer can bias a transistor in two ways: for current biasing the current through the device is imposed and the terminal voltages are the dependent variables; for voltage biasing the terminal voltages are imposed and the current is the dependent variable. For the sake of simplicity of the equations and calculations, the source and bulk are supposed connected so that $V_{SB} = 0$ and no bulk-effect occurs. However, if a bulk-effect does occur in the circuit, the extra mismatch due to the mismatch in the bulk-effect coefficients can in first order simply be considered as an extra degradation of the V_T matching of the transistors, so that most equations can still be used. Of course, for the optimization of the biasing, the dependence of the bulk-effect on the bias voltages should be taken into account and slightly different results will be obtained.

For voltage biasing the terminal voltages are imposed and the current I_{DS} is the dependent variable as is illustrated in figure 3.6(a); the current depends on the terminal voltages and on the transistor parameters:

$$I_{DS} = f(V_{GS}, V_{DS}, V_{T0}, \beta) \tag{3.16}$$

For a pair of transistors with an identical V_{GS}, the difference in their currents is calculated using (3.13):

$$\Delta I_{DS} = \frac{\partial I_{DS}}{\partial \beta}\Delta\beta + \frac{\partial I_{DS}}{\partial V_{T0}}\Delta V_{T0} \tag{3.17}$$

From the device equations in appendix A one can conclude for all regions of operation that:

$$\frac{\partial I_{DS}}{\partial V_{T0}} = \frac{\partial I_{DS}}{\partial V_{GS}} = -g_m \tag{3.18}$$

$$\frac{\partial I_{DS}}{\partial \beta} = \frac{I_{DS}}{\beta} \tag{3.19}$$

so that the relative current difference in all operating regions of the transistor is expressed by:

$$\left(\frac{\Delta I_{DS}}{I_{DS}}\right) = \left(\frac{\Delta\beta}{\beta}\right) - \left(\frac{g_m}{I_{DS}}\right) \Delta V_{T0} \tag{3.20}$$

and the variance of the current difference is:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \left(\frac{\sigma(\Delta\beta)}{\beta}\right)^2 + \left(\frac{g_m}{I_{DS}}\right)^2 \sigma^2(\Delta V_{T0}) \tag{3.21}$$

In a current biasing scheme as illustrated in figure 3.6(b) the current is imposed and the gate-source voltage V_{GS} is the dependent variable. Expressions for the calculation of the V_{GS} as a function of the current are not always available. However (3.16) can be rewritten as an implicit function for V_{GS}:

$$I_{DS} - f(V_{GS}, V_{DS}, V_{T0}, \beta) = 0 \tag{3.22}$$

and the partial derivatives of V_{GS} can be expressed as:

$$\frac{\partial V_{GS}}{\partial\beta} = -\left(\frac{\partial I_{DS}}{\partial\beta}\right)\left(\frac{\partial I_{DS}}{\partial V_{GS}}\right)^{(-1)} \tag{3.23}$$

$$= \frac{-I_{DS}}{\beta g_m} \tag{3.24}$$

$$\frac{\partial V_{GS}}{\partial V_{T0}} = -\left(\frac{\partial I_{DS}}{\partial V_{T0}}\right)\left(\frac{\partial I_{DS}}{\partial V_{GS}}\right)^{(-1)} \tag{3.25}$$

$$= 1 \tag{3.26}$$

so that the gate-source voltage difference and its variance become:

$$\Delta V_{GS} = \Delta V_{T0} - \left(\frac{I_{DS}}{g_m}\right)\left(\frac{\Delta\beta}{\beta}\right) \tag{3.27}$$

$$\sigma^2(\Delta V_{GS}) = \sigma^2(\Delta V_{T0}) + \left(\frac{I_{DS}}{g_m}\right)^2 \left(\frac{\sigma(\Delta\beta)}{\beta}\right)^2 \tag{3.28}$$

3.3.2.1 Strong inversion

For a transistor biased in strong inversion and in saturation, the g_m/I_{DS} is $2/(V_{GS} - V_T)$ (see (A.22)) so that (3.21) and (3.28) can be rewritten as:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \left(\frac{\sigma(\beta)}{\beta}\right)^2 + \frac{4\sigma^2(V_{T0})}{(V_{GS} - V_T)^2} \tag{3.29}$$

$$\sigma^2(\Delta V_{GS}) = \sigma^2(V_{T0}) + \frac{(V_{GS} - V_T)^2}{4}\left(\frac{\sigma(\beta)}{\beta}\right)^2 \tag{3.30}$$

Substituting the models of (3.1) and (3.3) in (3.29) and (3.30), we obtain for closely spaced devices:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \frac{1}{WL}\left(A_\beta^2 + \frac{4A_{VT0}^2}{(V_{GS}-V_T)^2}\right) \tag{3.31}$$

$$\sigma^2(\Delta V_{GS}) = \frac{1}{WL}\left(A_{VT0}^2 + \frac{(V_{GS}-V_T)^2 A_\beta^2}{4}\right) \tag{3.32}$$

The accuracy of the gate voltage or drain current is dependent on the bias point of the devices or $(V_{GS}-V_T)$ and on their gate-area. For a technology a *corner* gate-drive voltage $(V_{GS}-V_T)_m$ is defined for which the effect of the V_{T0} and β mismatch on the gate voltage or drain current is of equal size:

$$(V_{GS}-V_T)_m = 2A_{VT0}/A_\beta \tag{3.33}$$

In (3.31) and (3.32) we observe that for a circuit with a bias point with a $(V_{GS}-V_T)$ smaller than $(V_{GS}-V_T)_m$ the effect of the V_{T0} mismatch is dominant, whereas for a $(V_{GS}-V_T)$ larger than $(V_{GS}-V_T)_m$ the effect of the β mismatch dominates. In table 3.1 the values of $(V_{GS}-V_T)_m$ are listed for a few CMOS technologies. It is clear that in practical circuits the $(V_{GS}-V_T)$ will be smaller than $(V_{GS}-V_T)_m$ so that the V_{T0} mismatch is dominant over the β mismatch for the calculation of the accuracy of the circuit behavior. In practice, equations (3.31) and (3.32) can be approximated by:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 \approx \frac{4A_{VT0}^2}{WL(V_{GS}-V_T)^2} \tag{3.34}$$

$$\sigma^2(\Delta V_{GS}) \approx \frac{A_{VT0}^2}{WL} \tag{3.35}$$

The approximation error due to neglecting the β mismatch on the standard deviation σ for the above equations is equal to $((V_{GS}-V_T)/(V_{GS}-V_T)_m)^2/2$, and is small for typical transistor bias conditions; for a nMOS transistor in the 0.7 μm technology biased with a $(V_{GS}-V_T)$ of 0.2 V the error on σ is only 1 %.

3.3.2.2 Weak inversion

When the transistors are biased in weak inversion the g_m/I_{DS} is $1/(nU_T)$ (see (A.4)) so that (3.21) and (3.28) can be rewritten as:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \left(\frac{\sigma(\beta)}{\beta}\right)^2 + \frac{\sigma^2(V_{T0})}{(nU_T)^2} \tag{3.36}$$

$$\sigma^2(\Delta V_{GS}) = \sigma^2(V_{T0}) + (nU_T)^2\left(\frac{\sigma(\beta)}{\beta}\right)^2 \tag{3.37}$$

By substituting (3.1) and (3.3) in (3.36) and (3.37), the relative current variation and gate voltage variation of two closely spaced devices become:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 = \frac{1}{WL}\left(A_\beta^2 + \frac{A_{VT0}^2}{(nU_T)^2}\right) \tag{3.38}$$

$$\sigma^2(\Delta V_{GS}) = \frac{1}{WL}\left(A_{VT0}^2 + (nU_T)^2 A_\beta^2\right) \tag{3.39}$$

The weak-inversion slope parameter n typically has values from 1 to 2 and U_T is 25.8 mV at room temperature. From table 3.1 we can conclude that the V_{T0} mismatch dominates the accuracy calculations so that (3.38) and (3.39) can be simplified to:

$$\left(\frac{\sigma(\Delta I_{DS})}{I_{DS}}\right)^2 \approx \frac{A_{VT0}^2}{WL(nU_T)^2} \tag{3.40}$$

$$\sigma^2(\Delta V_{GS}) \approx \frac{A_{VT0}^2}{WL} \tag{3.41}$$

The approximation error due to neglecting the β mismatch on the standard deviation σ for the above equations is equal to $(2nU_T/(V_{GS} - V_T)_m)^2/2$; for a nMOS transistor in the 0.7 μm technology the error on the σ is only 0.1 % at room temperature.

3.4 Implications of transistor mismatch on the behavior and design of elementary stages

In the forthcoming sections the implications of transistor mismatch on the speed, power consumption and accuracy of elementary current and voltage processing stages is studied. This allows to draw guidelines for optimal design of these circuits. Furthermore it provides the background for a discussion of the implications of transistor mismatch on the performance of general analog VLSI systems [Kin 96d].

3.4.1 Current processing circuits

The current amplifier, represented in figure 3.7, is a basic current processing stage. The output transistor M_2 is a parallel connection of A unit-transistors of the same size as M_1, with a gate-width W and gate-length L, so that the current amplification factor is A.

The speed performance of the current amplifier is the highest frequency that can be processed by the current amplifier, and depends on the bandwidth. The bandwidth of this circuit is in first-order determined by the g_m of the input transistor M_1 and the parallel connection of the gate capacitors of both transistors. For transistors biased in strong inversion, the bandwidth of the amplifier is:

$$\begin{aligned} \mathrm{BW} &= \frac{g_{m1}}{2\pi(C_{GS1} + C_{GS2})} \\ &= \frac{3I_B}{2\pi(A+1)(V_{GS} - V_T)C_{ox}WL} \end{aligned} \tag{3.42}$$

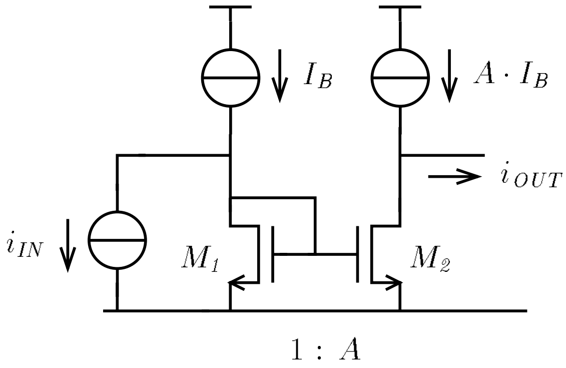


Figure 3.7: Basic current amplifier.

The DC power consumption of the amplifier is :

$$P = (A+1) I_B V_{DD} \tag{3.43}$$

The relative accuracy of the current processing is determined by the maximal input signal RMS value I_{inRMS} and the 3σ value of the input referred offset current I_{OS} and is defined as:

$$\mathrm{Acc}_{\mathrm{rel}} = \frac{I_{inRMS}}{3\sigma(I_{OS})} \tag{3.44}$$

By using the 3σ value of the offset current, the accuracy specification is met with a probability of about 99.7%. This probability that a circuit block meets its specifications, has a direct impact on the yield of the total chip or system. In complex systems with many stages, an even higher probability can be necessary to obtain a high yield and more than the 3σ has to be accounted for in (3.44).

Due to the effect of mismatches in the transistors, an error occurs in the current mirroring and for a zero input current a non-zero output current exists. The input referred offset current I_{OS} is by definition the current that has to be applied to the input to obtain a zero output current; it has to compensate for the variation in the current of M_1 and M_2. To calculate the I_{OS}, we first calculate the errors in the currents of M_1 and M_2. The standard deviation of the current in transistor M_1 and in a unit-transistor of M_2 is derived from (3.34):

$$\sigma(I_{UNIT}) = \frac{1}{\sqrt{2}} I_B \frac{2A_{VT0}}{(V_{GS} - V_T)\sqrt{WL}} \tag{3.45}$$

where the term $1/\sqrt{2}$ is necessary to calculate the variance of the parameter from the variance of the difference of parameters – see (3.15). The total current in M_2 is the sum of the currents of the individual unit transistors, which are statistically independent quantities, so that the standard deviation of the current of M_2 is $\sigma(I_{OUT}) = \sqrt{A}\sigma(I_{UNIT})$.

Since the current amplification is A, we can express the standard deviation of the input offset current as follows:

$$\begin{aligned}\sigma(I_{OS}) &= \sqrt{\frac{\sigma^2(I_{OUT})}{A^2} + \sigma^2(I_{UNIT})} \\ &= I_B \frac{\sqrt{2}A_{VT0}}{(V_{GS}-V_T)\sqrt{WL}}\sqrt{\frac{A+1}{A}}\end{aligned} \tag{3.46}$$

For a typical bias modulation index of 1/2 the I_{inRMS} is $I_B/(2\sqrt{2})$ and the relative accuracy ($\mathrm{Acc_{rel}}$) of the current processing then becomes:

$$\mathrm{Acc_{rel}} = \frac{\sqrt{WL}(V_{GS}-V_T)}{12A_{VT0}}\sqrt{\frac{A+1}{A}} \tag{3.47}$$

With the expressions for the different circuit performance parameters at hand, we can now develop a relation for the total performance or quality of the circuit design. The quality of the circuit is better if its bandwidth, gain and accuracy are large and its power consumption is low. When we consider the ratio *Speed·Accuracy²/Power* for the current amplifier, we obtain:

$$\frac{\mathrm{BW Acc_{rel}^2}}{P} = \frac{(V_{GS}-V_T)}{96\pi C_{ox} A_{VT0}^2 V_{DD}} \frac{A}{(A+1)^3} \tag{3.48}$$

which we can rewrite for large gains A as:

$$\frac{\mathrm{Gain^2 BW Acc_{rel}^2}}{P} = \frac{(V_{GS}-V_T)}{96\pi C_{ox} A_{VT0}^2 V_{DD}} \tag{3.49}$$

Very important conclusions for the design of a current amplification stage can be drawn from this relation:

- The *total* performance of the amplifier is only dependent on technology constants and on the chosen bias point of the stage and is independent of the transistor sizes. To obtain the best *total* performance, a current processing stage must be designed with a *large* $(V_{GS}-V_T)$. It is common knowledge that to improve the accuracy of a current mirror a large $(V_{GS}-V_T)$ has to be used [Lak 94]. From (3.29) one can indeed conclude that the accuracy performance of a current mirror is improved by increasing $(V_{GS}-V_T)$. On the other hand, increasing $(V_{GS}-V_T)$ reduces the g_m of the stage so that the speed performance is degraded and that a higher bias current must be used to obtain a certain speed, resulting in higher power consumption. However, (3.49) shows that increasing $(V_{GS}-V_T)$ will result in the best possible trade-off between speed, accuracy, gain and power consumption for a current processing stage !

For a current processing stage the $(V_{GS} - V_T)$ is typically limited to V_{DD} $/2$[a]. As a result, the optimal performance becomes:

$$\frac{\mathrm{Gain}^2\mathrm{BWAcc}^2_{\mathrm{rel}}}{P} = \frac{1}{192\pi C_{ox} A^2_{VT0}} \tag{3.50}$$

- The different performance specifications of a current amplifier are linked and depend on technological and physical constants only ! Equation (3.50) shows that a designer can only trade one specification for the other. Due to the impact of transistor mismatch, he cannot choose the different performance specifications independently ! When we rewrite (3.50) as:

  $$P = 192\pi C_{ox} A^2_{VT0}\mathrm{Gain}^2\mathrm{BWAcc}^2_{\mathrm{rel}} \tag{3.51}$$

  Transistor mismatch puts a boundary on the minimal power consumption by a current amplification stage for a given gain, speed and accuracy specification.

- An important limitation of current processing stages is the quadratic dependence of the power consumption on the gain of the amplifier in (3.51). The fundamental reason for the appearance of this term is related to the physics of the MOS transistor. By changing the gate voltage, the conductivity of the channel is controlled and thus the current is controlled. On the physical level, a MOS transistor acts as a voltage dependent current source or a transconductor. If we try use this device as a current amplifier, we basically operate the device in an unnatural way. The only way to make current gain is by a parallel connection of several transistors to the gate of a diode-connected transistor, that does the current to voltage conversion. The more gain we try to make, the more load we put on this gate for the same transconductance and the speed in (3.42) reduces. Moreover, the more gain, the larger the bias current in the output transistor and the larger the power consumption in (3.43).

  The accuracy in (3.47) on the other hand, is in first order independent of the realized gain. The input signal swing is only limited by the modulation index we can allow. The maximal modulation index depends on the distortion as is discussed in section 3.7. The bias current of the output stage scales with the gain so that the modulation index at the output is the same as at the input. As a result the gain does not influence the maximal signal swing or the relative accuracy.

  Due to the effect of the gain on the bandwidth and the power consumption, a quadratic dependence on the gain of the quality of the circuit in (3.49) and of the the minimal power consumption of the circuit in (3.51) emerges.

[a] for small feature sizes, the appearance of velocity saturation [Lak 94] can further reduce the maximal $(V_{GS} - V_T)$ that can be used under which the above derivations remain valid.

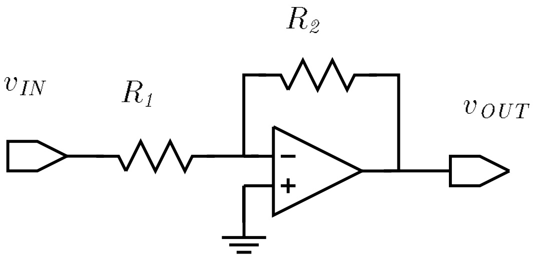


Figure 3.8: System schematic for a basic voltage amplification block.

Exact case

When the full expression for the variance of the relative current difference (3.31) is used, including both the V_{T0} and β mismatch, the following expression for the quality of a current amplifier with a large gain is derived:

$$\frac{\text{Gain}^2\text{BWAcc}^2_{\text{rel}}}{P} = \frac{1}{24\pi C_{ox} V_{DD}\left(A^2_\beta(V_{GS}-V_T)+\frac{4A^2_{VT0}}{(V_{GS}-V_T)}\right)} \tag{3.52}$$

The value of $(V_{GS}-V_T)$ can now be optimized to obtain a maximal total circuit performance: starting with a small $(V_{GS}-V_T)$ and increasing it, the term proportional to A_{VT0} in the denominator of (3.52) decreases and the quality improves; for large values of $(V_{GS}-V_T)$, however, the first term proportional to A_β increases and the quality decreases again. An optimal value for $(V_{GS}-V_T)$ exists and can easily be calculated: the optimum in circuit performance is reached for a $(V_{GS}-V_T)$ equal to $(V_{GS}-V_T)_m$- see (3.33), and we then obtain the best possible total circuit performance:

$$\frac{\text{Gain}^2\text{BWAcc}^2_{\text{rel}}}{P} = \frac{1}{96\pi C_{ox} V_{DD} A_\beta A_{VT0}} \tag{3.53}$$

In the previous section, taking only the effect of V_T offset into account, we concluded that for current processing circuits the $(V_{GS}-V_T)$ of the stage has to be maximized to optimize performance. However, by taking also the effect of β mismatch into account, we conclude that $(V_{GS}-V_T)$ should not be increased beyond $(V_{GS}-V_T)_m$. Furthermore, if the supply voltage allows it, a current processing stage should be biased with $(V_{GS}-V_T) = (V_{GS}-V_T)_m$ which yields the optimum in the trade-off between the different specifications.

3.4.2 Voltage processing circuits

The inverting voltage amplifier stage of figure 3.8 is a basic voltage processing stage. It consists of an operational amplifier (opamp) with a negative resistive feedback. The input voltage is converted into a current by R_1 and the virtual ground that is created at the negative input terminal of the opamp. This current is converted into the output voltage by R_2 so that the closed loop amplification of the system A_{CL} is determined by the ratio R_2 $/R_1$, which is well controlled over process variations. This behavior of the system is observed as long as the open-loop gain of the opamp is much larger than the closed loop gain.

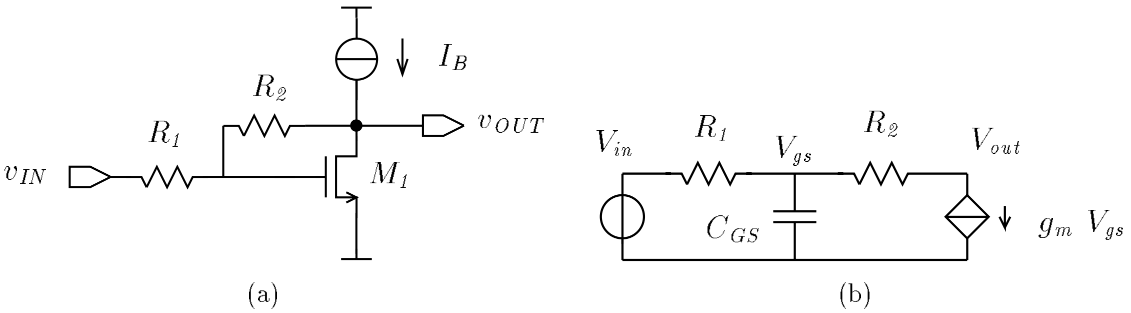


Figure 3.9: (a) One transistor voltage amplifier implementation; (b) small signal equivalent.

3.4.2.1 One transistor voltage amplifier

The most simple implementation for the operational amplifier in figure 3.8 is obtained with a single transistor as in figure 3.9(a). From the small signal equivalent in figure 3.9(b), we derive the frequency response of the system:

$$\begin{aligned} A(s) &= \frac{V_{out}}{V_{in}} \\ &= -\frac{R_2}{R_1}\left(\frac{1 - 1/(g_m R_2)}{1 + 1/(g_m R_1) + s/(g_m/C_{GS})}\right) \end{aligned} \tag{3.54}$$

We observe in (3.54) that in order to define the amplification accurately, the $1/g_m$ of the transistor has to be chosen considerably smaller than the resistors R_1 and R_2. The low-frequency closed-loop amplification of the stage is then $A_{CL} = R_2/R_1$ and the closed-loop frequency response is approximately given by:

$$A(s) \approx -\frac{R_2}{R_1}\left(\frac{1}{1 + s/(g_m/C_{GS})}\right) \tag{3.55}$$

so that we obtain the following relation for the the bandwidth of the amplifier:

$$\mathrm{BW} \approx g_m/(2\pi C_{GS}) \tag{3.56}$$

A) Strong inversion First we consider the performance of this voltage amplifier when the transistor is biased in strong inversion. Using the relations for the g_m and C_{GS} in strong inversion from appendix A, we can express the bandwidth of the amplifier as follows:

$$\mathrm{BW} = \frac{3 I_B}{2\pi (V_{GS} - V_T) W L C_{ox}} \tag{3.57}$$

The DC power consumption of the circuit is:

$$P = V_{DD} I_B \tag{3.58}$$

The relative accuracy ($\mathrm{Acc_{rel}}$) of the circuit is determined by the offset voltage of the opamp (V_{OS}) with respect to the maximal RMS value of the input signal (V_{inRMS}). The maximal voltage swing at the output is in first order: $V_{outPP} = V_{DD}$, so that $V_{inRMS} = V_{DD}/(2\sqrt{2}\text{Gain})$. The standard deviation of the offset voltage of the one transistor opamp is determined by equation (3.35), but from (3.15) we conclude that the offset voltage is $\sqrt{2}$ times smaller since the variation on V_{GS} is important. The expression for the relative accuracy can now be evaluated as:

$$\text{Acc}_{\text{rel}} = \frac{V_{inRMS}}{3\sigma(V_{OS})} \tag{3.59}$$

$$= \frac{V_{DD}\sqrt{WL}}{6A_{VT0}\text{Gain}} \tag{3.60}$$

To obtain a better insight in the possible trade-offs between the different circuit parameters, we consider again the total performance of the voltage amplifier. Combining (3.57), (3.58) and (3.60), we obtain:

$$\frac{\text{Gain}^2\text{BW}\text{Acc}^2_{\text{rel}}}{P} = \frac{V_{DD}}{24\pi(V_{GS}-V_T)A^2_{VT0}C_{ox}} \tag{3.61}$$

From (3.61) some very important lessons can be learned for the design of a voltage amplifier:

- The *total* performance of a voltage amplifier is maximized by *lowering* $(V_{GS}-V_T)$. As far as speed requirements allow it the stage should be biased with the lowest possible $(V_{GS}-V_T)$. Again it is common knowledge that to achieve a low offset voltage a small $(V_{GS}-V_T)$ has to be used as can also be derived from (3.32) [Lak 94]. However, (3.61) shows that not only a good accuracy but the best trade-off between speed, gain, accuracy and power consumption is obtained for a small $(V_{GS}-V_T)$. A typical minimal value of $(V_{GS}-V_T) = 0.2\ V$ is derived from device physics - see Appendix A. The best attainable performance in strong inversion is:

$$\frac{\text{Gain}^2\text{BW}\text{Acc}^2_{\text{rel}}}{P} = \frac{5V_{DD}}{24\pi A^2_{VT0}C_{ox}} \tag{3.62}$$

- Also for voltage designs the different performance specifications are linked by physical and technological constants only ! The designer can - once the optimal bias is chosen - not optimize the different specifications independently but can only trade one specification for an other. When we rewrite (3.62), we obtain that the minimal required power consumption is fixed for a given gain, speed and accuracy by the impact of transistor mismatch:

$$P = \frac{24\pi}{5V_{DD}}A^2_{VT0}C_{ox}\text{Gain}^2\text{BW}\text{Acc}^2_{\text{rel}} \tag{3.63}$$

- In (3.63) we observe again a quadratic dependence of the power consumption on the gain of the voltage amplifier. In a voltage amplifier nor the power consumption, nor the bandwidth are in first order dependent on the realized gain. However, when the gain is increased, the maximal input signal reduces proportionally since the maximal output swing is limited by the supply voltage. This limits the maximal input signal swing; consequently, the maximal attainable accuracy for a given power supply voltage reduces proportional to the gain (3.60) and the power consumption increases quadratically with the gain (3.63).

B) Weak inversion In the previous section it was shown that the *total performance* of a voltage amplifier improves for smaller $(V_{GS} - V_T)$. Consequently, the best total performance is obtained when the transistor is biased in weak-inversion.

The relations for the power consumption (3.58) and for the relative accuracy (3.60) remain valid. Since source and bulk are connected, the capacitance that determines the speed performance in weak inversion is the gate-bulk capacitance, which is:

$$C_{GB} = \frac{(n-1)}{n} C_{ox} W L \tag{3.64}$$

With the expression for g_m in weak inversion, we derive the following relation for the bandwidth of the amplifier operating in weak-inversion:

$$\mathrm{BW} = \frac{g_m}{2\pi C_{GB}} = \frac{I_B}{2\pi U_T (n-1) C_{ox} W L} \tag{3.65}$$

And we can now calculate the total performance of the amplifier in weak inversion as:

$$\frac{\mathrm{Gain}^2 \mathrm{BW} \mathrm{Acc}^2_{\mathrm{rel}}}{P} = \frac{V_{DD}}{72\pi (n-1) U_T A^2_{VT0} C_{ox}} \tag{3.66}$$

For the 0.7 μm technology e.g. the technology constants are $V_{DD} = 5\ V$, $C_{ox} = 2\ fF/\mu m^2$, $A_{VT0} = 13\ mV\mu m$, and n is approximately 1.5 and $U_T = 25.8\ mV$ at room temperature; the optimal quality in weak inversion calculated with (3.66) is then 5 times better than the best quality in strong inversion calculated with (3.62). This implies that for the same gain, bandwidth and accuracy specifications 5 times less power is required for an amplifier operated in weak inversion.

3.4.2.2 Differential pair voltage amplifier

A simple differential implementation of the operational amplifier in figure 3.8 is a differential pair as is shown in figure 3.10. Similarly as for the one transistor amplifier the $1/g_m$ of the input transistors is designed larger than the value of the resistors resistors R_1 and R_2 to obtain a well controlled closed-loop gain. As a result the bandwidth of the amplifier in a high precision design is limited by the pole caused by the input capacitance and it is easily calculated that: $\mathrm{BW} = g_m/(2\pi C_{GS})$.

The power consumption of the circuit is $P = 2V_{DD}I$. The relative accuracy $(\mathrm{Acc}_{\mathrm{rel}})$ of the circuit is determined by the offset voltage of the opamp, or the V_{T0} mismatch of

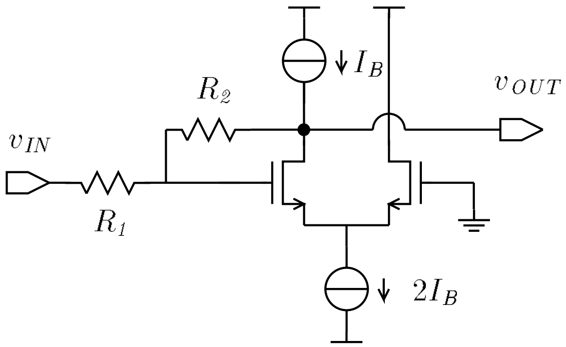


Figure 3.10: Basic voltage amplification stage implemented with a differential pair amplifier.

the input transistors, in respect to the maximal RMS of the input signal, which is in first order $V_{DD}/(2\sqrt{2}\ \text{Gain})$:

$$
\begin{aligned}
\text{Acc}_{\text{rel}} &= \frac{V_{DD}}{6\sqrt{2}\ \sigma(V_{OS})\ \text{Gain}} \\
&= \frac{V_{DD}\sqrt{WL}}{6\sqrt{2}\ A_{VT0}\ \text{Gain}}
\end{aligned} \tag{3.67}
$$

For transistors biased in strong inversion and saturation, the quality of the design is again expressed as:

$$
\frac{\text{Gain}^2\text{BW}\text{Acc}_{\text{rel}}^2}{P} = \frac{V_{DD}}{96\pi(V_{GS}-V_T)A_{VT0}^2 C_{ox}} \tag{3.68}
$$

The *total* performance of this voltage processing stage is again maximized by *lowering* $(V_{GS}-V_T)$ towards operation in weak inversion. This result shows that the differential implementation consumes about 4 times more power for the same gain, bandwidth and accuracy specifications as the single transistor implementation (see equation (3.61)). The power consumption is doubled since two transistors require two times the bias current and their offset voltage is $\sqrt{2}$ as for the single transistor (see equation (3.15)); both effects result in 4 fold increase of the power consumption.

Summary

To obtain a low power consumption, voltage circuits must be operated in weak inversion as much as possible. However, the maximal attainable frequency in weak inversion is limited, since the maximal current in weak inversion is also limited. Weak inversion operation is only possible for relatively low speed applications. In section 3.4.4.2 the limitations on the use of (3.61) and (3.66) to calculate the necessary power consumption are studied in detail. A second practical, but important, limitation of the use of weak inversion, is the limited availability from industrial chip foundries of good model parameters for the simulation of transistors in weak inversion.

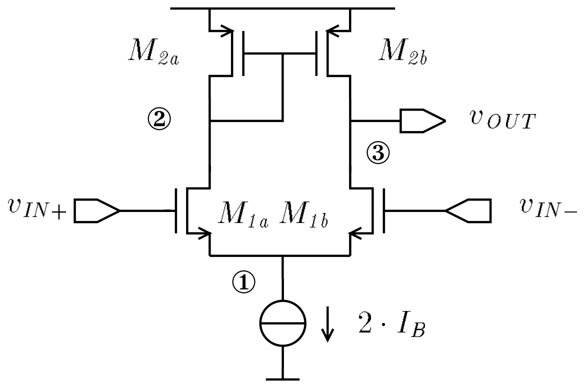


Figure 3.11: Load compensated OTA.

Up to now we have discussed the mismatch limitation on the trade-offs between the different performance parameters for simple signal processing stages. This makes the analytical analysis straightforward and closed-form expressions are obtained. We will now proceed with the analysis of a more complex circuit implementation of a load compensated OTA. However, increasing the circuit complexity, increases the degrees of freedom in the design and involves the design of several stages; in the calculations some reasonable assumptions and approximations are made to obtain again closed-form expressions.

3.4.2.3 Load compensated OTA voltage processing stage

The operational amplifier that is used in the basic voltage amplifier of figure 3.8 can also be implemented with the operational transconductance amplifier represented in figure 3.11. It is the basic schematic for many more sofisticated load compensated operational amplifiers. As such, its design and the trade-off between its specifications, is representative for the design procedure of many multi-stage operational amplifiers. The circuit operates as follows: transistors M_{1a-b} transform the differential component of the input voltages into a differential current; they act as a voltage to current converter. The signal current is converted to a single ended output voltage by the current mirror M_{2a-b} and the output conductance of M_{1b} and M_{2b}.

These amplifiers are typically used in a feedback configuration like e.g. the configuration of figure 3.8. The stability of the feedback system is then first concern of the circuit designer; an amplifier is only useful if its stability is guaranteed. Due to the high complexity of the circuit, its transfer function is of higher order. Basically, a pole is associated with each circuit node. The presence of a second and higher order poles in the open-loop transfer function implies that the open-loop frequency response of the OTA has to be adapted in order to obtain a safe phase and gain margin for all possible feedback configurations. The maximal speed or frequency performance that can be attained is then determined by the frequency of the second pole in the open-loop transfer function. The gain-bandwidth (GBW) of the amplifier must be made K_{stab} times smaller than the second pole and K_{stab} is at least 2 for a phase margin better than 60 degrees.

The second-pole (f_2) in the open-loop transfer function is located at node ②, the gates of transistors M_{2a} and M_{2b}, and is calculated as:

$$\begin{aligned} f_2 &= \frac{g_{m2}}{2\pi(C_{GS2a}+C_{GS2b})} \\ &= \frac{3I_B}{4\pi C_{ox} W_2 L_2 (V_{GS}-V_T)_2} \end{aligned} \tag{3.69}$$

so that the maximal GBW of the amplifier becomes:

$$\begin{aligned} \mathrm{GBW} &= \frac{f_2}{K_{stab}} \\ &= \frac{3I_B}{4\pi K_{stab} C_{ox} W_2 L_2 (V_{GS}-V_T)_2} \end{aligned} \tag{3.70}$$

The accuracy of the amplifier depends on the equivalent input referred offset voltage (V_{OS}). The V_{OS} is determined by the voltage matching of the transistor pair M_{1a-b} and the accuracy of the current mirror M_{2a-b}. For the calculation of V_{OS}, the internal voltage gain of the amplifier A_{in} from the input to gate of M_{2a-b} is an important factor. It determines the attenuation of the influence of the errors in the second stage on the input referred offset. In order to simplify the expressions we assume that all devices have equal length ($L_1 = L_2$), which is a reasonable assumption for high speed designs where all signal transistors are designed with a minimal length to obtain a maximal transconductance and minimal node and input capacitances. All transistors have the same bias current, so that the A_{in} is expressed by:

$$A_{in} = \frac{g_{m1}}{g_{m2}} = \frac{(V_{GS}-V_T)_2}{(V_{GS}-V_T)_1} = \sqrt{\frac{W_1}{W_2}} \tag{3.71}$$

The standard deviation of the offset voltage is calculated as follows:

$$\sigma^2(V_{OS}) = \sigma^2(V_{T01}) + \left(\frac{\sigma(V_{T02})}{A_{in}}\right)^2 \tag{3.72}$$

$$= \frac{A^2_{VT0n}}{W_1 L_1} + \frac{A^2_{VT0p}}{W_2 L_2 A^2_{in}} \tag{3.73}$$

$$= \frac{1}{W_2 L_2 A^2_{in}} \left(A^2_{VT0n} + A^2_{VT0p}\right) \tag{3.74}$$

Equation (3.72) shows that by increasing the internal gain A_{in}, the effect of the mismatch in the current mirror on the input referred offset is lowered. However, from (3.71) we conclude that the gain can only be increased by decreasing the width W_2 and consequently by decreasing the area of the current mirror transistors since the transistors have the same lengths; the increase in gain goes at the cost of a reduction of the matching; therefor the nMOS as well as the pMOS matching, expressed by respectively A_{VT0n} and A_{VT0p}, are equally important in the final expression for the offset voltage (3.74).

Using the expression (3.74), the relative accuracy ($\mathrm{Acc}_{\mathrm{rel}}$) is calculated from (3.59). The DC power consumption of the amplifier is given by:

$$P = 2I_B V_{DD} \tag{3.75}$$

The quality of the circuit implementation is evaluated from:

$$\frac{\mathrm{Gain}^2 \mathrm{GBW} \mathrm{Acc}^2_{\mathrm{rel}}}{P} = \frac{V_{DD}}{192\pi K_{stab} C_{ox}(A^2_{VT0n} + A^2_{VT0p})} \cdot \left(\frac{A_{in}}{(V_{GS} - V_T)_1} \right) \tag{3.76}$$

where Gain is the gain in the closed loop configuration determined by the applied feedback.

The speed performance of the total voltage processing circuit, as depicted in figure 3.8, with an OTA is determined by the maximal frequency for which the loop gain is larger than 1. Since the GBW of open-loop transfer function of the OTA is fixed, we obtain the following expression for the BW of the total amplifier:

$$\mathrm{BW} = \frac{\mathrm{GBW}}{Gain} \tag{3.77}$$

Substituting this result in (3.76), we derive the following relation for the voltage processing system:

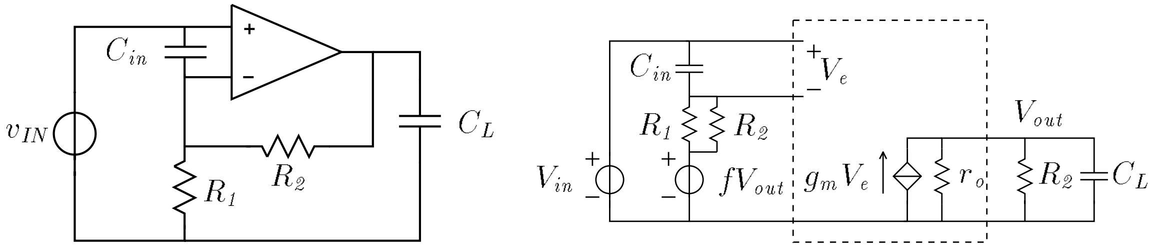
$$\frac{\mathrm{Gain}^3 \mathrm{BW} \mathrm{Acc}^2_{\mathrm{rel}}}{P} = \frac{V_{DD}}{192\pi K_{stab} C_{ox}(A^2_{VT0n} + A^2_{VT0p})} \cdot \left(\frac{A_{in}}{(V_{GS} - V_T)_1} \right) \tag{3.78}$$

This expression leads to the following conclusions:

- The quality of the design improves by biasing the input transistors M_{1a-b}, which operate in *voltage mode*, with a *low* ($V_{GS} - V_T$).

- By choosing a high internal gain and consequently by biasing the transistors M_{2a-b}, which operate in *current mode*, with a *high* ($V_{GS} - V_T$) better performance is obtained. The choice of the gain in multi-stage amplifiers is further examined in section 3.5.1.

- The minimal power to obtain a given speed and accuracy depends on the *cubic* of the gain of the voltage processing block. As in the one transistor implementation a second order dependence is due to the limited output swing of the amplifier, which results in a maximal input signal and relative accuracy inversely proportional to the gain. However an extra gain dependence is due to the presence of a second pole in the amplifier; the power consumption is proportional to the GBW of the open-loop but in closed loop the bandwidth is the gain times smaller. Due to the higher complexity which brings extra constraints relating to the stability into the design, more power is consumed.



(a) (b)

Figure 3.12: (a) OTA based voltage processing system schematic; (b) small signal equivalent where $f = R_1/(R_1 + R_2)$.

3.4.3 Feedback systems and general OTA design

In the previous sections we showed that the *Speed·Accuracy²/Power* product of the basic voltage and current circuit blocks is fixed by technological constants. This result is obtained because in the simple blocks the same transistor is responsible for the accuracy, speed and power specification of the circuit. The bandwidth limiting capacitor, for instance, is inversely proportional to the offset so that the *Speed·Accuracy²* product becomes independent of the transistor sizing. In this section we will further generalize these findings to general analog systems; we study in detail how the different performance specifications are related in a general feedback voltage processing system, implemented with an OTA. Similar results can be obtained for other types of systems.

In figure 3.12(a) a schematic for a voltage processing system is shown. The output voltage is fed back in series to the input so that a high input impedance is obtained, which is desirable in voltage processing systems [Gra 84, Lak 94]. The power consumption of this system is determined by the current consumption of the OTA; the bias current of the OTA is proportional to the required g_{mIN} of the input stage, which is on its turn dependent on the wanted gain-bandwidth product of the system. General feedback theory and general opamp/OTA theory [Lak 94, Gra 84] show that in a system as in figure 3.12(a), the gain in closed loop A_{CL} times the bandwidth in closed loop BW_{CL} is a constant called the gain-bandwidth product GBW. An expression for the GBW of many types of OTA's is [Lak 94, Gra 84]:

$$\mathrm{GBW} = A_{CL}\mathrm{BW}_{CL} = \frac{g_{mIN}}{2\pi C_d} \tag{3.79}$$

where C_d is the capacitor that is associated with the dominant pole of the open-loop transfer function of the OTA and g_{mIN} is the transconductance of the input stage. If we assume an input stage operating in strong inversion and we derive the following expression for the power consumption by using (3.58) and (3.79):

$$P = \pi A_{CL}\mathrm{BW}_{CL}C_dV_{DD}(V_{GS} - V_T) \tag{3.80}$$

The relative accuracy of the system is expressed by (3.59). The offset voltage of the OTA is in first order determined by the area of the input devices and consequently is related to the input capacitance as is shown in (3.101) derived in section 3.5.1. The output swing is limited to V_{DD} so that the maximal input signal is $V_{inRMS} = V_{DD}/(2\sqrt{2}A_{CL})$ and the relative accuracy is then given by:

$$\mathrm{Acc}^2_{\mathrm{rel}} = \frac{{V_{DD}}^2 C_{in}}{24 A^2_{CL} C_{ox} A^2_{VT0}} \tag{3.81}$$

and we can rewrite the power consumption as:

$$P = 24\pi \frac{(V_{GS}-V_T)}{V_{DD}} C_{ox} A^2_{VT0} A^3_{CL} \mathrm{BW}_{CL} \mathrm{Acc}^2_{\mathrm{rel}} \left(\frac{C_d}{C_{in}}\right) \tag{3.82}$$

We obtain a relationship between the specifications of a general voltage processing system implemented with an OTA very similar to the relationship obtained for the basic voltage processing stage in (3.61). If a relation between the the input capacitance C_{in} and the capacitance of the dominant pole C_d exists, the trade-off between the different circuit specifications is, also for the general system, determined by technological and physical constants only.

When the stability of the system in figure 3.12(a) is studied it becomes apparent that in any OTA design the fraction C_d/C_{in} will be larger than or equal to 1. From the small signal equivalent for an implementation with an OTA in figure 3.12(b), the loop transfer function is derived:

$$T(s) = \frac{g_m R_1}{(1 + sR_2C_L)(1 + sC_{in}R_1)} \tag{3.83}$$

using the following approximations:

$$A_{CL} = \left(\frac{R_2}{R_1} + 1\right) > 1 \text{ or } R_2 > R_1 \tag{3.84}$$

$$\text{and} \qquad r_o \gg R_2 \tag{3.85}$$

Due to the presence of two capacitors in the system, we obtain two poles[b]:

$$f_a = \frac{1}{2\pi R_1 C_{in}} \tag{3.86}$$

$$f_b = \frac{1}{2\pi R_2 C_L} \tag{3.87}$$

To evaluate the position of these poles we need an estimation of the relative magnitude of the C_{in} versus C_L. The load capacitance of the voltage processing stage is determined

[b] the voltage input source resistance is assumed to be zero: $R_s = 0$; a non-zero R_s is in series with R_1 for the calculation of the pole f_a. In (3.86) we have thus assumed that $R_1 \gg R_s$ and the location of the pole can be chosen by the designer. In the case that $R_s \gg R_1$, the location of f_a is fixed by R_s and the input capacitance C_{in}, so that no design freedom exists for the position of f_a. That case is covered in section 3.5.1 and equation (3.104).

by the input capacitance of the next stage. As is explained in section 3.5.1, the input capacitance of the second stage C_L will be smaller as the C_{in}, since the offset requirements for the second stage will be less severe. In order to meet a certain accuracy specification in a multi-stage system, we have to divide the allowed total offset over the different stages as can be concluded from (3.97); (3.101) indicates that the ratio of C_L/C_{in} will be proportional to $\sigma^2(V_{os1})/\sigma^2(V_{os2})$. A good design compromise is to make $\sigma(V_{os1})/\sigma(V_{os2}) = 1/\sqrt{A_{CL}}$, since then the offset of the total system is dominated by the offset of the first stage only. When we approximate the exact expression for A_{CL} in (3.84) by R_2/R_1, we derive the following relative position of the two poles:

$$f_b = \frac{f_a}{\sqrt{A_{CL}}} \tag{3.88}$$

so that f_b is the dominant pole and that C_L plays the role of C_d in (3.82). To assure a stable feed-back system the second pole $f_{nd} = f_a$ must be at least larger or equal than the unity-gain frequency of the loop-transfer function T(s) to have a phase-margin of at least 45 degrees so that we obtain the following condition:

$$\begin{aligned} \frac{f_{nd}}{\mathrm{GBW}} &\geq 1 \\ \frac{1}{2\pi R_1 C_{in}} \frac{2\pi C_L}{g_m} &\geq 1 \\ \frac{C_L}{C_{in}} &\geq g_m R_1 = T_{@DC} \overset{!}{\geq} 1 \end{aligned} \tag{3.89}$$

To guarantee stability the ratio C_L/C_{in} must be made larger as $g_m R_1$ which is the loop gain at low frequencies as can be concluded from (3.83) and which is always at least larger than 1. So to attain stability, we have to increase the load capacitance C_L since we cannot make C_{in} smaller due to the accuracy specification.

Due to stability requirements, the ratio of capacitors in (3.82) will always be larger than 1. Also for other design choices of $\sigma(V_{os1})/\sigma(V_{os2})$, stability requirements impose that $C_d/C_{in} \geq 1$. We can conclude that the power consumption of a voltage feedback system is thus larger than:

$$P \geq 24\pi \frac{(V_{GS} - V_T)}{V_{DD}} C_{ox} A_{VT0}^2 A_{CL}^3 \mathrm{BW}_{CL} \mathrm{Acc}_{\mathrm{rel}}^2 \tag{3.90}$$

Equation (3.90) leads to the following conclusions:

- the minimal power consumption is limited by the effect of transistor mismatch and the quality of the technology is expressed by $C_{ox} A_{VT0}^2$; and also the maximal attainable *Speed·Accuracy²/Power* product is limited by the input transistor;
- to optimize the total combined performance of a voltage processing system the input stages have to biased with a low $(V_{GS} - V_T)$ or in weak inversion;

- in open-loop stages[c] the minimal power consumption is proportional to the square of the Gain in (3.51), (3.63) and (3.66); when feedback is used the minimal power is dependent on the *cubic* of the closed loop gain (Gain) in (3.78) and A_{CL} in (3.90) due to the effect of the limited gain-bandwidth of an opamp or OTA and the stability requirements.

3.4.4 Circuit design guidelines

We have studied several basic circuit stages for current or voltage signal processing. In this section we summarize the conclusions of the different designs and we highlight the limitations of the applicability of the different trade-off relationships.

3.4.4.1 Current processing stages

From the design of the current amplifier it can be concluded that in order to obtain optimal *total performance* i.e. high speed, high accuracy and low power consumption, a current processing stage must be biased with a *high* $(V_{GS} - V_T)$ as long as V_T mismatches are dominant. However the $(V_{GS} - V_T)$ of the stage should not be increased above the $(V_{GS} - V_T)_m$ of the technology. The overall best performance is obtained when the current processing stage is biased at $(V_{GS} - V_T)_m$. In practical situations this will be impossible due to other specifications like, for instance, signal swing and limited power supply voltage.

3.4.4.2 Voltage processing stages

A voltage processing stage must be biased with a *low gate-overdrive voltage* $(V_{GS} - V_T)$ in order to obtain an optimal *total performance*; the lowest power consumption is obtained in weak-inversion. However the maximal attainable intrinsic speed in a transistor decreases for small $(V_{GS} - V_T)$; so the minimal $(V_{GS} - V_T)$ can be imposed by the required speed performance. In the next paragraph we illustrate this for the one-transistor voltage amplifier of section 3.4.2.1.

Limitations on the choice of $(V_{GS} - V_T)$**:** Lowering the $(V_{GS} - V_T)$ of a transistor in strong inversion, lowers its maximal cut-off frequency f_T which is defined as [Lak 94]:

$$f_T = \frac{g_m}{2\pi C_{GS}} = \frac{3}{4\pi}\frac{\mu(V_{GS} - V_T)}{L^2} \qquad (3.91)$$

Since the bandwidth and thus the maximal operating frequency of a voltage processing stage is proportional to the f_T of the transistors, see (3.56), lowering the $(V_{GS} - V_T)$ limits the maximal operating frequency of the circuit. The minimal gate-overdrive voltage $(V_{GS} - V_T)$, which guarantees a strong inversion behavior of the transistor is typically

[c] although the simple processing stage in section 3.4.2.1 has feedback, its performance trade-offs are those of an open-loop stage since no second pole is taken into account in the calculations.

0.2 Volts (appendix A). We define the maximal frequency that can be processed for a $(V_{GS} - V_T)$ of 0.2 Volts, as the corner frequency f_{coII}:

$$f_{coII} = f_T \mid_{(V_{GS}-V_T)=0.2} = \frac{3}{20\pi}\frac{\mu}{L^2} \tag{3.92}$$

As long as the operating frequency is lower than the corner frequency f_{coII} and a minimal gate length is used, equation (3.63) for the calculation of the power consumption is valid.

For frequencies higher than f_{coII}, however, the $(V_{GS} - V_T)$ must be made proportional to the required bandwidth. When we substitute this result into (3.61) the minimal power consumption for the one transistor amplifier satisfies:

$$P = \frac{8\pi^2}{3} C_{ox} A_{VT0}^2 \frac{L^2}{\mu V_{DD}} \text{Acc}_{\text{rel}}^2 \text{BW}^2 \text{Gain}^2 \tag{3.93}$$

The power consumption is now proportional to the square of the operating frequency !

Equation (3.91) suggests that the absolute maximal frequency response that can be achieved in a MOS technology is limited by the minimal length L and the maximal $(V_{GS} - V_T)$ that can be used. But the maximal speed of electrons in silicon is limited to v_{sat} $(\approx 10^5 m/s)$ so that at high $(V_{GS} - V_T)$ biases, the transistor goes in velocity saturation [Lak 94]. The voltage current relation is not quadratic in $(V_{GS} - V_T)$ anymore; the transconductance g_m of a transistor with fixed dimensions does not increase with the current anymore so that the maximal cut-off frequency f_{comax} becomes a constant, independent of biasing:

$$f_{comax} = \frac{1}{2\pi}\frac{v_{sat}}{L} \tag{3.94}$$

This is the absolute maximal frequency that can be achieved in a technology.

In section 3.4.2.1 the lowest power consumption is obtained for a transistor biased in the weak-inversion or sub-threshold regime. The maximal bias current I_{Mwi} that still biases the devices in weak-inversion is limited to [Lak 94]:

$$I_{Mwi} = 0.2\mu n C_{ox} \frac{W}{L} U_T^2 \tag{3.95}$$

The maximal cut-off frequency in weak inversion is then derived by substituting (3.95) into (3.65) and is given by:

$$f_{coI} = \frac{n\mu U_T}{10\pi(n-1)L^2} \tag{3.96}$$

so that for frequency requirements below f_{coI}, the transistor can be biased in weak-inversion and the power consumption is limited by (3.66) for a given accuracy and gain.

Using the technology parameters from appendix A, we calculate the values of the different cut-off frequencies for the 0.7 μm CMOS technology; the maximal frequency f_{coI}

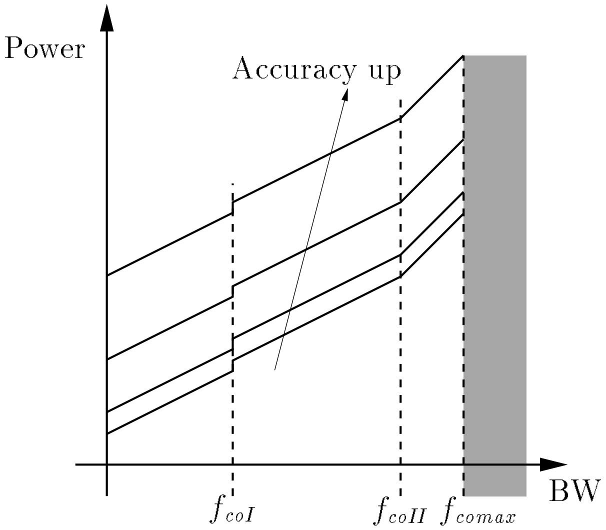


Figure 3.13: Evolution of the minimal power consumption as a function of the required bandwidth for the circuit, for different accuracies.

in weak-inversion is 240 MHz; the f_{coII} lies at 4.6 GHz, and the absolute maximal frequency f_{comax} is 22 GHz. These numbers are very optimistic however, since in the first order analysis we have only taken into account the loading by the gate-source capacitances C_{GS}. In practical circuits, extra loads due to the e.g. the parasitic drain-bulk capacitors and interconnect or biasing sources parasitic capacitors reduce the frequency response of the circuits. Moreover, the derived expressions are for an open-loop amplifier, whereas in closed loop systems extra margins of safety have to be considered to assure system stability so that the maximal achievable bandwidth further reduces. Since the dominant parasitic load capacitors are typically also proportional to the transistor sizing (see discussion of α_{DB} in appendix A), the qualitative evolution of the power consumption discussed in this section remains valid for more complex circuits.

When the minimal power consumption for a one transistor voltage amplifier is plotted as a function of the required bandwidth BW, four operating regions can be distinguished, as is illustrated in figure 3.13:

- for BW $\leq f_{coI}$, the transistor can be biased in weak inversion and the power consumption is proportional to BW;
- for $f_{coI} <$ BW $\leq f_{coII}$, the transistor can be biased in strong inversion with the minimal $(V_{GS} - V_T)$ of 0.2 Volts and the power consumption is proportional to BW but is the proportionality constant is about 5 times higher than in weak-inversion as derived in section 3.4.2.1;
- for $f_{coII} <$ BW $\leq f_{comax}$, the transistor must be operated in strong inversion with a $(V_{GS} - V_T)$ proportional to the BW so that the power consumption scales with the square of BW;

- a BW $> f_{comax}$ is not achievable in the technology due to velocity saturation effects in silicon.

For all regions the power consumption scales with the square of the accuracy and the square of the gain. The discrete jump in the power consumption at f_{coI} is due to the limited accuracy of the transistor models we have used in the derivation of the minimal power consumption. We have considered a fixed limit between strong inversion and weak-inversion operation; however a smooth transition region called moderate inversion [Tsi 88] exists between the two regions so that also a smooth transition is observed in practical circuits for the power consumption; unfortunately simple and accurate hand-calculation models for the transistor behavior in this region are not available.

Limiting the value of $(V_{GS} - V_T)$ to obtain a minimal power consumption is thus limited by the required speed performance of the circuit block. The more demanding the speed requirements are and the more they approach the technology limits, the more excessive the power consumption becomes. This conclusion can be generalized for all voltage processing stages. In section 3.4.3, for instance, the bandwidth for a general OTA design in a feedback configuration is given by (3.79) and determined by the transconductance of the input stage and the dominant capacitor. As is shown, the dominant capacitor is larger than or equal to the input capacitance for stability reasons. The speed performance of the stage is thus limited by the intrinsic speed of the input device. The optimal power consumption is achieved for minimal $(V_{GS} - V_T)$ in the input stage, but also in that design reducing $(V_{GS} - V_T)$ can only be tolerated as long as the bandwidth specification allows it. The power consumption as a function of the required speed performance will have a similar characteristic as in figure 3.13.

3.5 Implications of mismatch on analog system performance

In the previous sections we have developed expressions for the trade-offs between the different circuit performance specifications of simple building blocks by studying their detailed design equations. From (3.51), (3.63) and (3.78) it is clear that a minimal power consumption is imposed for a given operation frequency (or bandwidth) and a given precision by the impact of device mismatch. This minimal power consumption is proportional to the matching quality of the technology which is expressed by $C_{ox}A^2_{VT0}$. In systems containing many building blocks the relations between the different specifications are more complex since more degrees of freedom exist. There is extra room for circuit and topology optimization but the transistor mismatch has basically the same impact. In this section we will derive relations for the power consumption as a function of the other specifications from basic general design equations. We show that a general relation can be derived for the minimal power consumption of a signal processing system due to the effect of the mismatch in the components.

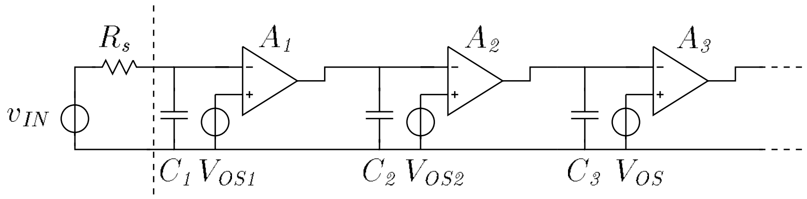


Figure 3.14: Schematic representation for a multi-stage voltage processing system.

3.5.1 General Multi-stage voltage designs

A general multi-stage voltage processing system is represented in figure 3.14. For a current signal processing system or a system architecture with currents and voltages intermixed similar schematics can be drawn and similar relationships can be derived. In 3.14 the offset voltage, input capacitance and gain of each stage is indicated. The relative accuracy - see (3.59) - of the total system is determined by the input signal RMS value and the equivalent input referred offset voltage which is calculated as follows:

$$\sigma(V_{OSeq}) = \sqrt{\sigma^2(V_{OS1}) + (\frac{\sigma(V_{OS2})}{A_1})^2 + (\frac{\sigma(V_{OS3})}{A_1 A_2})^2 + \ldots} \tag{3.97}$$

Mismatch effects put a lower boundary on the smallest signal that can be processed in a system. As a result, their influence is most important at the input stage where the signal levels are small. Therefore a designer will try to make the largest possible gain in the first stage A_1 so that the influence of the following stages on the accuracy specification is negligible[d]. The expression for $\sigma(V_{OSeq})$ is dominated by the term of the offset of the first stage so that the relative accuracy is approximately:

$$\mathrm{Acc}_{\mathrm{rel}} = \frac{V_{inRMS}}{\sigma(V_{OSeq})} \approx \frac{V_{inRMS}}{3\sigma(V_{OS1})} \tag{3.98}$$

The offset voltage of a stage is inversely proportional to the area of the input devices; the input capacitance is proportional to the area of the same input devices; thus a general relation between the input capacitance and the offset voltage exists. If we assume an input stage with a single transistor biased in strong inversion and since V_T mismatch is dominant, we obtain the following expressions:

$$C_{in} = 2/3C_{ox}WL \tag{3.99}$$

$$\sigma^2(V_{OS}) = \frac{A_{VT0}}{2WL} \tag{3.100}$$

$$C_{in} = \frac{C_{ox}A_{VT0}^2}{3\sigma^2(V_{OS})} \tag{3.101}$$

[d] Since the power consumption of a block increases with the gain (see e.g. (3.63)), whereas the common required supply voltage for all blocks is determined by the wanted input swing and total gain, it is most power efficient to use as few stages as possible with the largest gain possible.

For a stage with a differential input, the offset voltage is $\sqrt{2}$ times larger but the input capacitance is 2 times smaller so that the relation (3.101) between input capacitance and offset voltage remains valid. By applying (3.98) and (3.101), the input capacitance of the multi-stage system becomes:

$$C_1 \approx \frac{C_{ox} A_{VT0}^2}{3\sigma^2(V_{OS_{eq}})} \tag{3.102}$$

The maximal speed of the system is of course determined by the bandwidth of the amplification stages; however, the pole generated by the input capacitor C_1 and the source resistance R_s is the upper boundary for the system speed:

$$\begin{aligned} f_{lim} &= \frac{1}{2\pi R_s C_1} \\ f_{lim} &= 3/2 \frac{{V_{os1}}^2}{2\pi R_s C_{ox} A_{VT0}^2} \end{aligned} \tag{3.103}$$

From (3.98) it can be concluded that:

$$f_{lim} \mathrm{Acc}_{\mathrm{rel}}^2 \approx \frac{{V_{inRMS}}^2}{6\pi R_s C_{ox} A_{VT0}^2} \tag{3.104}$$

For the design of a multi-stage system the following conclusions are important:

- In high accuracy systems the best matching specifications have to be achieved in the *first or input stages* where the signal levels are the smallest; and the largest possible signal amplification has to be done as soon as possible.
- In a general design, (3.104) proves that the speed and the accuracy of the system are interdependent for a given input source impedance and their combination cannot be chosen freely but is dependent on the matching quality of the technology.

In this section we assumed that the input signal source drives the input capacitance of the signal processing system directly; then the power required for the signal processing is delivered by the input signal source. Many practical input voltage signal sources have relatively high output impedances, so that from (3.104) it is clear that an unbuffered input is not possible for high speed systems. The system schematic has to be changed so that the input capacitance is buffered for the input signal source and so that a high input impedance is obtained over a wide frequency range. Due to this extra buffering the power required to drive the input capacitance, is now delivered by the first block of the signal processing system itself, which we will study in the next section.

3.5.2 Limit imposed by mismatch on minimal power consumption

Mismatch is a random process and and its effect on the circuit behavior can be translated into time-invariant random DC error signals called offset voltages or currents. The random

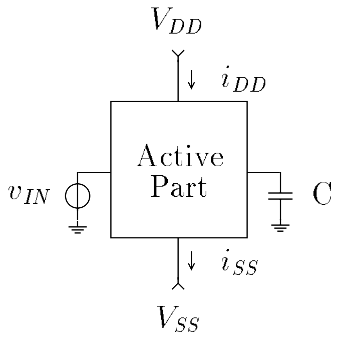


Figure 3.15: General schematic for an active circuit driving a capacitive load.

variations or the variance of the offset signals will be smaller if large devices are used since an averaging and smoothing occurs of the spatial errors sources responsible for the device mismatch. A certain accuracy will thus require a certain device area. However, this will lead to unavoidable parasitic capacitors at the input of the circuit proportional to the device area and thus inversely proportional to the offset signals as is demonstrated in (3.101).

When we want to buffer a certain voltage signal $v_S = V_s sin(2\pi ft)$ and drive it across a capacitor C, a current has to be delivered by the active part proportional to the voltage swing and the conductance of the capacitor at the signal frequency f: $i_C = V_s 2\pi fCcos(\omega t)$. In figure 3.15 a general schematic is represented for an active part driving a capacitor. Although no power is dissipated by the capacitor since its current and voltage have a 90 degrees phase-shift, for all active parts operating from a practical power voltage supply, power will be dissipated in the active part. The power consumption of the active part will depend on its efficiency which is a function of the mode of operation.

For a *class A operation* a DC bias is added to all signals equal to their amplitude. The minimum required supply voltage for the active part is $V_{DD} - V_{SS} = 2V_s$, whereas the minimum required DC bias current is $I_B = 2V_s\omega C$ so that the following relation for the power consumption in class A is obtained:

$$P = 16\pi fCV_{sRMS}{}^2 \tag{3.105}$$

When an active part operating in *class B* is used, no DC offsets are used but the active devices draw the signal current from the V_{DD} and deliver it into the capacitive load and they return the current from the load to the V_{SS}. The current signals i_{DD} from the V_{DD} supply and i_{SS} to the V_{SS} supply are depicted in figure 3.16. The average DC current flowing from the positive to the negative supply is $I_{DC} = (1/\pi)(2\pi fCV_s)$ so that the power consumption in class B is given by:

$$P = 8fCV_{sRMS}{}^2 \tag{3.106}$$

Other classes of operating modes exist for the implementation of active systems, like class C, E, F, S; they have much higher efficiencies but can only be used for very specific types of signals and applications whereas class A and B are almost generally applicable. From (3.105) and (3.106) we can conclude that due to the presence of capacitors in the system a

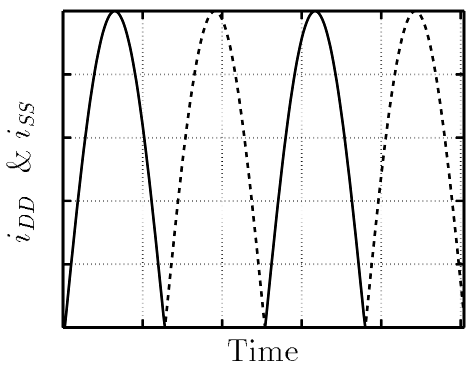


Figure 3.16: The supply currents for a class B active circuit driven with a sinusoidal input signal: (–) i_{DD} and (– –) i_{SS}.

certain amount of power will be dissipated in the system to perform the signal processing at a certain speed or frequency f.

At this point we have all the relations available to estimate the power consumption of a system for a certain signal processing operation. From the definition of the dynamic range of a building block as the largest signal that it can process over the smallest signal, we have $DR = V_{sRMS}/(3\sigma(V_{OS}))$ and using (3.101) and (3.106) we obtain the following expression for the power consumption:

$$P = 24C_{ox}A_{VT0}^2 fDR^2 \tag{3.107}$$

This relation leads to the following conclusion: an analog signal processing system will at least consume the power expressed in (3.107) due to the effect of mismatch to perform a signal processing operation at a given frequency f and with a certain accuracy or DR. Equation (3.107) is the most general expression of the fact that the *Speed·Accuracy²/Power* ratio is fixed by the technology mismatch quality.

3.5.3 Noise vs mismatch limits on minimal power consumption

Thermal noise and mismatch are both random processes and put a limit on the smallest signal that can be processed in a circuit; both physical phenomena will impose a minimal power consumption to achieve a certain DR specification and speed. In this section the minimal power consumption due to noise is derived and the limits imposed by mismatch and imposed by thermal noise are compared.

In [Vit 90b] the effects of thermal noise on the power consumption of a circuit is studied, which we repeat here for completeness. At a node with a capacitance C and driven by an impedance R, the total integrated thermal noise is:

$${V_{nRMS}}^2 = \frac{kT}{C} \tag{3.108}$$

with k the Boltzmann constant and T the absolute temperature. The dynamic range of the building block is given by $DR = V_{sRMS}/V_{nRMS}$, so that the minimal power consumption to

Technology	Type	Mismatch	Noise
		$24C_{ox}A^2_{VT0}$	$8kT$
		[fJ]	[fJ]
$2.5\mu m$ [Pel 89]	nMOS pMOS	4.3e-2 -	3.3e-5 3.3e-5
$1.2\mu m$ [Bas 95]	nMOS pMOS	2.12e-2 -	3.3e-5 3.3e-5
$0.7\mu m$	nMOS pMOS	6.3e-3 -	3.3e-5 3.3e-5

Table 3.4: The minimal energy per cycle imposed by mismatch and by noise for a dynamic range of 1.

achieve a certain DR imposed by thermal noise is given by:

$$P = 8kTfDR^2 \tag{3.109}$$

To be able to compare these fundamental limits to the performance of realized circuits which all have different operating frequencies, not the power consumption but the energy per cycle P/f is evaluated as a function of the DR. In figure 3.17 the noise limit and mismatch limit for the minimal energy consumption are plotted as a function of the DR. The mismatch limit is technology dependent, since the product $C_{ox}A^2_{VT0}$ is a technology parameter and not a physical constant like kT. For present-day sub-micron CMOS technologies the limit on power consumption imposed by mismatch is about two orders of magnitude more important than the limit imposed by thermal noise, which can also be concluded from the values in table 3.4.

Analog Filters In a high order analog filter the minimal power consumption per pole of the filter by the impact of noise is given by (3.109). In figure 3.17 the high dynamic range analog filter circuits referenced in [Vit 90b] are also represented. The dynamic range of filters is in first order not sensitive to matching or offset voltages. Mismatch will mainly influence the accuracy of the filter coefficients, the distortion performance and the power supply noise or common mode rejection characteristics of the filter which do not directly influence the dynamic range. So filter realizations only optimized to low power and high dynamic range or distortion, can consume less than predicted (3.107).

High speed A/D converters, are a better benchmark because offsets or mismatch limits the bit accuracy directly. So the minimal power consumption of A/D converters is clearly limited by mismatch. In figure 3.17 the energy per cycle of several A/D converters is represented. The high accuracy architectures have typically lower speeds and include digital or analog error correction circuitry, so they are only about 2 orders of magnitude from optimal performance. The lower bit converters use very high speed architectures and

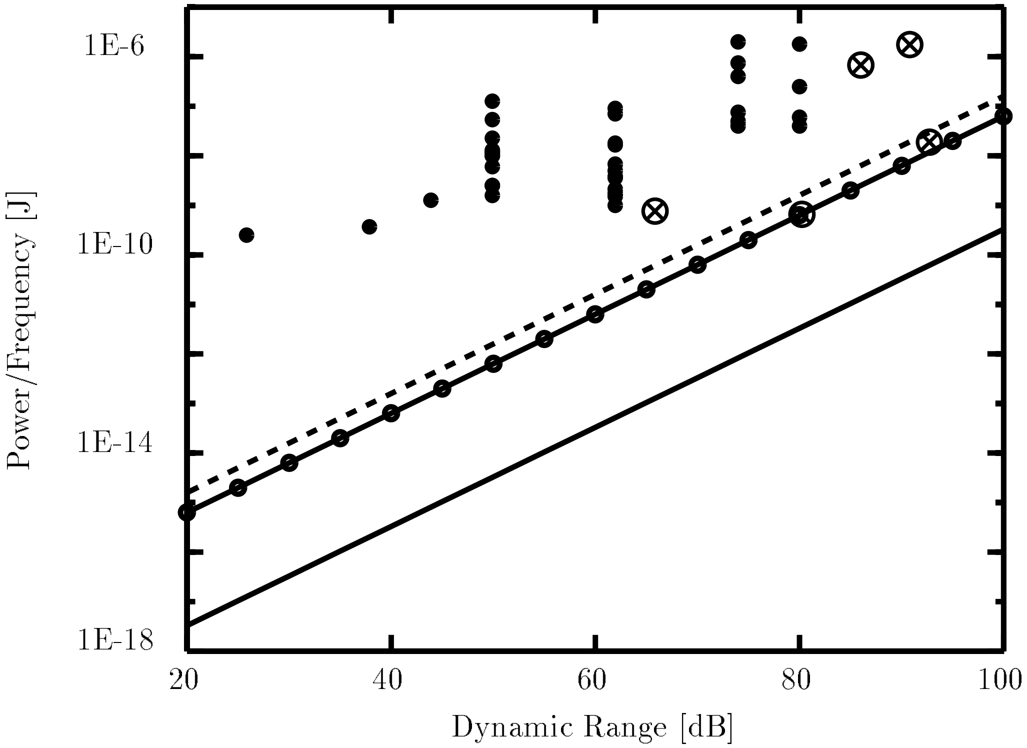


Figure 3.17: Comparison of the impact of thermal noise and mismatching on the power consumption of analog systems: (–) noise limit; mismatch limits: (– –) 1.2 μm and (–o–) 0.7 μm CMOS; Realized Filters (⊗) [Vit 90b] and A/D converters (•).

cannot rely on error correction so their performance is 2 to 4 orders from the optimal. The limit performance takes only the consumption of the input stage into account but in practical circuits the subsequent stages also consume considerable amounts of power. The large discrepancies are further also due to the extra power taken up by parasitics.

The derived performance limit caused by mismatch is of course only valid for converter architectures for which the accuracy relies on component matching. Many converter architectures, like e.g. $\Delta - \Sigma$ or algorithmic converters, exchange conversion speed for accuracy and are made insensitive to component matching at the cost of lower conversion speeds. Their performance is typically limited by noise [Dij 94]. Also in [Pel 94], the impact of transistor mismatch on the power consumption of high-speed A/D converters is discussed. The analysis is based on a different calculation path but the predicted minimal power consumption is of the same order of magnitude as our results.

Summary From the results in this section we conclude that in high speed, high accuracy analog systems, the effect of transistor mismatch imposes a higher minimal power consumption than the effect of thermal noise on the circuit specifications. It is important, however, to clearly understand the assumptions that lead to this conclusion.

Thermal noise is a fundamental physical limit to the minimal signal energy that can be used in electrical signal processing systems that arises from statistical thermodynamics [Mei 95]. Combining this limit with the best possible efficiency of a circuit, leads to the limit of (3.109). This limit cannot be broken by any circuit unless a more efficient architecture becomes available to transfer energy to a capacitive load from a DC power supply or other types of power supplies become available.

The origins of transistor mismatch on the other hand, are linked to the device structure and device physics and to the fabrication technology of integrated circuits; device mismatch originates from the stochastic nature of physical processes used for the fabrication like ion-implantations, diffusions or etching; the device structure using a channel in a doped material and its operation by modulating the channel resistance result in random fluctuations of the device properties and operation. For integrated circuit technologies as we use and fabricate them today, these physical limitations are very fundamental and device mismatch is unavoidable. In this perspective, the limits imposed by device mismatch, are restricted to signal processing systems realized using integrated circuits.[e] As such they are of course very important in the quest for minimal power consumption in integrated circuits but are of no importance for the fundamental physical limits of information processing.

As was demonstrated in this section, the mismatch limit is even more important than the noise limit in high speed analog systems. For high speed signal processing systems realized in present-day CMOS technologies the power consumption will be rather limited by the implications of device mismatch on the design and sizing of circuits to achieve a

[e] In other integrated circuit technologies then CMOS like bipolar Si or GaAs, the same or similar fabrication principles are used. Fluctuations in the device operation or device mismatches also occurs and are governed by similar relations as those described in this chapter so that we can generalize this conclusion to all integrated circuit technologies.

certain speed and accuracy specification than by the impact of thermal noise. However, when the system architecture or the application domain allows it, device mismatch can be circumvented by using offset compensating schematics as is discussed in section 3.6. As such the mismatch imposed limit is not important for all integrated circuits but unfortunately for a large part of high performance circuits.

3.5.4 Scaling of mismatch with technology feature size

3.5.4.1 Scaling of mismatch power limit

The minimal energy per cycle imposed by mismatch is proportional to $C_{ox}A_{VT0}^2$, as can be concluded from (3.107). The oxide capacitance per unit area is determined by the permittivity of silicondioxide ($\epsilon_m athitox$) and the oxide thickness (t_{ox}):

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \tag{3.110}$$

For smaller feature size technologies proportionally thinner gate-oxides are used so that C_{ox} is technology dependent.

Variations in the fixed oxide charge, silicon-silicondioxide surface-states charge and the depletion charge are the main causes for threshold voltage mismatch [Pel 89, Miz 94, Laks86, Shy 84]. These variations in charge are transformed in threshold voltage variations by the gate capacitance. Therefore the threshold voltage proportionality constant A_{VT0} is expected to be proportional to the gate-oxide thickness t_{ox}. Experiments where the oxide thickness is varied and all other technology parameters are kept constant, are reported in [Miz 94]; the V_{T0} mismatch is indeed proportional to the gate oxide thickness and for a zero gate oxide thickness a zero mismatch is extrapolated.

In figure 3.18 the A_{VT0} of several processes with different feature sizes is plotted versus the gate-oxide thickness t_{ox} [f]; a linear relation is indeed observed between A_{VT0} and t_{ox} for technologies with a feature size down to 0.5 μm. A linear least-squares fit yields a slope of 0.5 $mV\mu m/nm$ and a non-zero intercept of 3.4 $mV\mu m$. The physical origin of the non-zero intercept is not yet clear. However the gate-oxide thickness is not the only technological parameter that changes with the feature size. For smaller line-widths and thinner gate-oxides larger substrate doping levels (N_a) have to be used to avoid punch-through; larger doping levels increase the variance in the depletion charge and result in higher mismatching. A_{VT0} then becomes proportional to $t_{ox}N_a^{1/4}$ [Miz 94]. For a constant field scaling rule, the gate oxide thickness scales with $1/K$ and the substrate doping level scales with K so that the A_{VT0} scales with $1/K^{3/4}$. If the depletion charge mismatch is the dominant cause for threshold mismatch, and a constant field scaling can be assumed for the evolution of CMOS technologies, A_{VT0} is then expected to scale proportional to $t_{ox}^{3/4}$ when different processes are compared as in figure 3.18 and not linearly. This could be a possible explanation for the non-zero intercept when a linear fitting is erroneously

[f] I would like to thank Dr. M. Pelgrom of Philips Research in Eindhoven (NL.) for making available his measurement results reported in [Pel 89], which are also included in figure 3.18.

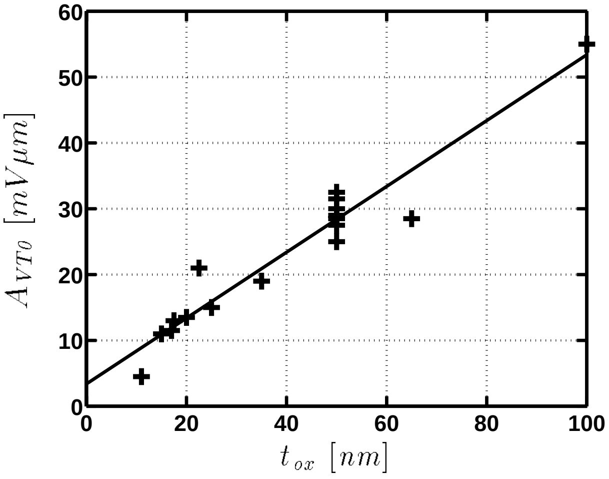


Figure 3.18: The experimental value of the threshold voltage mismatch proportionality constant A_{VT0} versus the gate-oxide thickness for different processes.

assumed. However other explanations, justifying a non-zero intercept have also been proposed [Pel 89]. It is clear however that in the presented figure many parameters change from technology to technology and many different scaling laws are in use which could all have an impact on the A_{VT0}.

From the previous paragraph we can conclude that to evaluate the scaling of the technology mismatch $C_{ox}A_{VT0}^2$ we can assume an almost linear dependence of A_{VT0} on t_{ox}; the technology mismatch $C_{ox}A_{VT0}^2$ thus scales with t_{ox} and the matching behavior of transistors improves for smaller feature sizes. This is also confirmed by the experimental data in table 3.1. For the analog systems, a migration towards smaller line-widths should improve the circuit performance.

However, the maximal supply voltage also reduces for smaller line-widths [Mea 94, Hu 93], so that smaller signal levels have to be used. Especially in voltage processing systems, this results in a deterioration of the performance (see also [Pel 96]). If we examine equations (3.62) and (3.66), we conclude that the quality of the circuit will decrease if the supply voltage is reduced. Consequently the scaling advantage for the quality of voltage designs with smaller technology line-widths is reduced. In figure 3.19 the quality of submicron technologies, normalized to the quality of the 0.7 μm technology, is plotted versus the technology line-width using the process data from [Hu 93] and appendix A; the A_{VT0} is approximated as 0.5 $mV\mu m/nm \cdot t_{ox}$. The –o– line uses $1/(C_{ox} \cdot A_{VT0}^2)$ as quality measure whereas the –x– line uses $V_{DD}/(C_{ox} \cdot A_{VT0}^2)$. The reduction of the supply voltage clearly reduces the scaling advantage of voltage processing stages.

For a current processing stage, the maximal signal swing is much less dependent on the available supply voltage. The voltage swing at the input depends only on the chosen bias modulation index. However, since the threshold voltage cannot be scaled proportionally to the supply voltage in order to limit the transistor cut-off leakage currents, the maximal $(V_{GS} - V_T)$ that can be used in a current processing stage reduces for lower supply voltages and the circuit performance will also degrade. However, in first order current processing

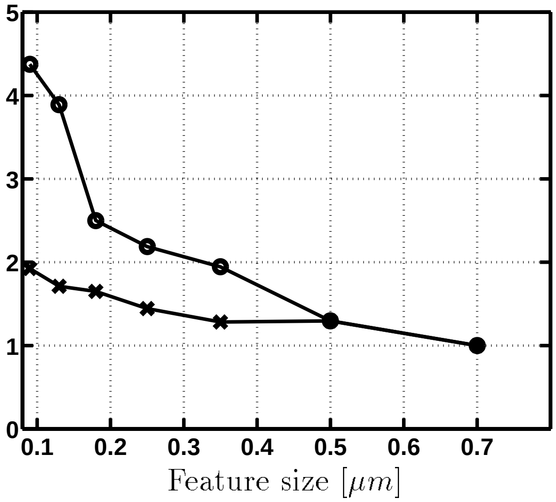


Figure 3.19: Scaling of the technology threshold matching quality as a function of the feature size; —o— does not take into account the decrease of the supply voltage for down-scaled technologies, whereas —x— accounts for the reduced quality of voltage processing stages due to the reduced available voltage swing in down-scaled technologies; the matching qualities are normalized to the value for the 0.7 μm technology.

stages are less sensitive to power supply voltage scaling.

Short channel effects also have a negative impact on the matching behavior as was demonstrated for the V_{T0} matching in section 3.2.3.E. Moreover, the increasing substrate doping levels in deeper sub-micron technologies make the parasitic drain to bulk and source to bulk capacitors relatively more and more important compared to the gate-oxide capacitance. This results in extra capacitive loading of the signal nodes and requires extra power to attain high speed operation.

We conclude that although the intrinsic matching quality of the technology improves for sub-micron and deep-sub-micron technologies, practical limitations make the theoretical boundary harder to achieve.

3.5.4.2 Relative importance of current factor and threshold voltage mismatches

At the start of the mismatch analysis we compared the relative importance of threshold voltage V_{T0} and current factor β mismatches on the behavior of transistors; for present-day processes the impact of the V_{T0} mismatch was clearly dominant. In the previous section the linear dependence of the A_{VT0} on the gate-oxide thickness was introduced so that the V_{T0} mismatch decreases for deeper sub-micron processes.

The proportionality constant A_β for the current factor has no clear relation to process parameters. The variation in β can be due to edge roughness, variations in the oxide thickness and mobility variations. The clear linear relation of the β mismatch with the gate-area excludes edge roughness as a dominant cause [Pel 89, Bas 95]; the low correlation

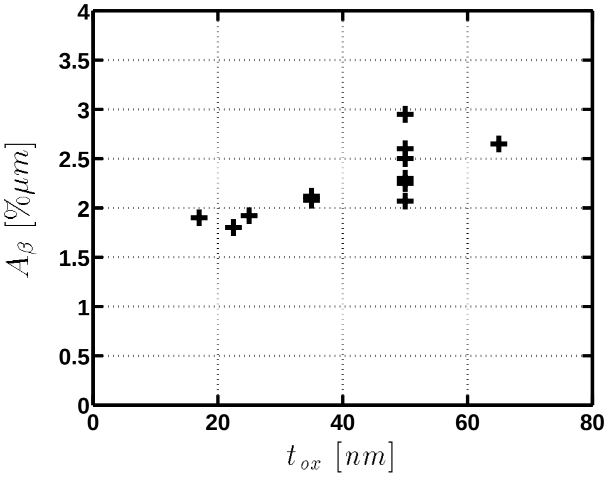


Figure 3.20: The experimental value of the current factor mismatch proportionality constant A_β versus the gate-oxide thickness for different processes.

with the V_{T0} mismatch excludes oxide variations as the dominant cause so that local mobility variations are the most probable dominant factor for β mismatches. In figure 3.20 the proportionality constants A_β are plotted for different processes[g] and we can conclude that the A_β is almost constant from technology to technology and has an approximate constant value of 2%μm.

When the scaling trends for A_{VT0} and A_β are compared, it is evident that the β mismatch gains in importance for deeper sub-micron technologies. This trend is confirmed by the values of the corner gate-overdrive voltage $(V_{GS}-V_T)_m$ in table 3.1 for different feature size technologies. For a slope in A_{VT0} with t_{ox} of 0.5 $mV\mu m/nm$ and a constant value of A_β of 2%μm a value of 200 mV is reached for $(V_{GS}-V_T)_m$ for a gate-oxide thickness of 4 nm; technologies with a feature size of about 0.25 μm are expected to use this t_{ox} [Mea 94, Hu 93]. For this technology the β mismatch will be at least as important and even more important as the V_{T0} mismatch for the calculation of the accuracy of circuits in the whole strong inversion region.

The circuit design guidelines derived in section 3.4.4 are no more valid at that point. When β mismatch is dominant, current mirrors have to biased with the smallest possible $(V_{GS}-V_T)$ to obtain optimal total performance. But current mirrors can probably be biased at the optimal $(V_{GS}-V_T)$ of $(V_{GS}-V_T)_m$- see section 3.4.1 - thanks to the low value of $(V_{GS}-V_T)_m$. At that bias point the minimal power consumption of current processing stages for a given speed and accuracy is proportional to $C_{ox}A_{VT0}A_\beta$ from (3.52); this indicates that a further scaling of the technology would not further improve the performance but would result in a constant performance as far as the assumptions of a constant A_β and an A_{VT0} proportional to t_{ox} remain valid. The total performance of voltage processing stages as in (3.61) becomes proportional to the cubic of $(V_{GS}-V_T)$ when β mismatch

[g] I would like to thank Dr. M. Pelgrom of Philips Research in Eindhoven (NL.) for making available his measurement results reported in [Pel 89], which are also included in figure 3.20.

is dominant over V_{T0} mismatch so that also the smallest possible $(V_{GS} - V_T)$ should be used. The minimal power consumption of voltage circuits then becomes proportional to $C_{ox}A_\beta^2$; this indicates that a further reduction of the oxide thickness would result in a worse performance for voltage processing circuits.

3.5.4.3 Summary

Thanks to the reduction of the V_{T0} mismatch with the oxide thickness, the move towards deeper sub-micron technologies will improve the intrinsic matching quality of the technology and should make lower power consumption possible. However, up to now no clear scaling of the β mismatch has occurred, so that for deep-sub-micron technologies of 0.25 μm and below, the current factor mismatch will become dominant; at that point further scaling of the technology is thought not to improve anymore the total performance of analog systems.

3.6 Techniques to reduce impact of mismatch

Many techniques have been developed to reduce the impact of mismatches on the performance of analog building blocks. The time-invariant nature of the offset signals caused by mismatches, allows the sampling of the error signals and subtracting them from the output. However, due to the limited availability of only accurate very short term analog memories, this compensation has to be repeated at regular time intervals during the system operation.

Auto-zero In comparators using an auto-zero compensation technique, the offset voltage of the comparator is first sampled in a dead period and stored dynamically on a capacitor; this voltage is then subtracted from the input voltage before the comparison. A simple implementation of this scheme and the timing of the control signals is represented in figure 3.21 [All 87, Lak 94]. Similar compensation techniques have been used in A/D converters and switched capacitor circuits to improve the accuracy.

However several limitations exist for the applicability of this technique. First of all, dead periods in the system operation have to be available so that the analog building block can be disconnected from the signal processing chain and its offset can be sampled. Not all system architectures have dead periods available.

After the sampling of the offset, it has to be stored on an analog memory. For integrated circuits, storing an analog signal as a charge quantity on a capacitor is the most practical technique. Every analog memory only has a limited retention time due to unavoidable leakage currents. Especially in high accuracy applications, where a highly accurate storage is required, the retention time becomes short so that the memory has to be refreshed with short intervals; the continuous operation period of the system becomes small for high accuracy signal processing.

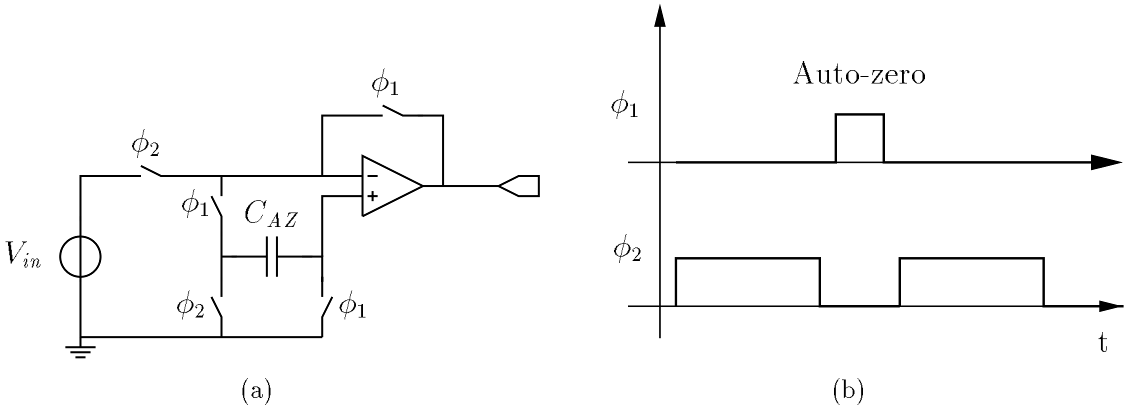


Figure 3.21: Auto-zero comparator schematic (a) to eliminate the effect of DC offset and (b) the timing for the switches.

The switching of the capacitor in 3.21, which is necessary to reconfigure the system back into an operational mode, introduces additional errors due to clock-feed-through and charge injection [VPg 88, Vit 90c]. Via the overlap capacitors between the hold capacitor and clock lines, an error charge is put on the holding capacitor when the clock signals change. These errors can be compensated by using complementary switches and complementary clock signals but they increase the system complexity even more. When the switch-transistors are turned off, the mobile carriers that made up the conductive channel flow away to the source and drain so that a charge is injected onto the hold-capacitor; this introduces an error in the sampled offset voltage. Extra compensation switches can be used to reduce the charge-injection but again increase the complexity of the system [VPg 86]. The error reduction is limited by the matching between the switches and the compensation switches. Again transistor mismatch puts a boundary on the possible accuracy, but now it impact will be at least on order of magnitude less important.

Chopping Chopper stabilization is another common technique for the reduction of the effect of offset voltages and 1/f noise on the accuracy of the signal processing [All 87]. The band-limited input signal of the amplifier is chopped by a multiplier driven with a 1/-1 signal, so that in the frequency domain it is modulated around the odd harmonics of the chopping frequency. This modulated signal is then amplified. The output signal is then again chopped so that is demodulated. The offset errors at the input of the amplifier are unaffected by the chopping and cause offsets in the output signal. The output chopping modulates these offset signals to high frequencies so that their effect in the passband is very small.

This techniques strongly reduces the effect of mismatch and 1/f noise. It requires however a high frequency operation of the amplifiers and the multipliers whereas the signal passband is only limited. Therefore, it is practically only useful for the design of high precision low frequency applications. In practical implementations a differential implementation of this effect is preferable in order to reduce the effects of power supply noise, charge injection and clock feed-through.

Trimming Trimming techniques are typically applied after the fabrication of the integrated system. The accuracy of the system is evaluated and is corrected by changing certain component parameters. Resistor values can be changed with laser-trimming. Discrete trimming can be achieved by switching extra resistors or capacitors in parallel to trim the accuracy. This can be achieved with digitally controlled switches. When the calibration is done once, long term digital memory cells have to be available which require for instance the availability of EEPROM devices in the technology, which is far from standard. When a self-calibration scheme is applied, the component will calibrate itself at every power-up. However this requires on-chip references so that the system accuracy can be measured; the accuracy of the references must be larger than the required system accuracy so that their design can be very complex. On the other hand analog fuses can be used for a one-time calibration but they are also not standard components and increase the technology cost. The accuracy improvement will be dependent on the resolution of the switch-able elements which can lead to important extra area consumption. The most important disadvantage of off-line trimming techniques is the extra testing or calibrating time required after fabrication which increases the cost of the device substantially.

The implication of the offset compensation on the power consumption of the building blocks can be very important. Basically if ideal compensation is available, mismatch has not to be taken into account in the design and the mismatch limitation on the minimal power consumption is not important. In practice however, the accuracy specifications are relaxed and only the last bits or percents are achieved with compensation or trimming. This implies that these techniques reduce the limit imposed by mismatch typically by an order of magnitude. When the area and complexity increase of the system can be tolerated or the fabrication technology has extra trimming facilities, offset compensation techniques should be applied and will enable important power saving.

3.7 Link with Harmonic distortion

Throughout this chapter we have focused on four main specifications of circuits and systems: accuracy, bandwidth or speed, gain and power consumption. However in many analog systems the linearity of the circuit or its distortion performance is of prime importance. In filter circuits, high linearity is necessary to avoid inter-modulation of unwanted signals into the passband or modulation of signal components. In all building blocks of telecommunication systems the distortion specifications play an important role in the design and performance trade-offs. To satisfy the linearity specification, the largest signal level in the circuit has to be reduced. Distortion has a direct impact on the maximal signals and thus influences the relative accuracy and dynamic range. Larger bias over signal ratio's have to be used so that more power is dissipated. In fact, when very high linearity specifications are to be met, the power consumption will be mainly originate from the impact of non-linearities [Wam 96b, Wam 96a].

In all calculations and derivations of this chapter we have made implicit assumptions about the allowed distortion. In the basic current amplifier (section 3.4.1) we assumed a bias current modulation index of 1/2 which has direct implications on the linearity of the amplifier. In the voltage processing circuits (section 3.4.2), the limitation on the maximal input signal was only limited by the maximal output swing; when linearity is important other constraints will exist and only a lower maximal input signal will be acceptable.

Since the distortion specification limits the maximal allowed signal, it has an impact on the relative accuracy of a circuit and consequently influences the quality of a circuit design. In voltage processing circuits, for instance, the relative accuracy is determined by the maximal input signal V_{inRMS} over the input referred offset signal V_{OS} (see also (3.59)):

$$\mathrm{Acc}_{\mathrm{rel}} = \frac{V_{inRMS}}{3\sigma(V_{OS})} \tag{3.111}$$

The maximal input signal depends on the allowed output signal swing and the gain (Gain) of the block. The output distortion reduces the allowed swing at the output to a fraction of the maximal swing, which is typically half the power supply $V_{DD}/2$; similarly the input distortion can further reduce the allowed maximal input signal swing. The linearity specifications and the distortion thus reduce the V_{inRMS} to:

$$V_{inRMS} = \alpha_{disto} \cdot \frac{V_{DD}}{2\sqrt{2}\ \mathrm{Gain}} \tag{3.112}$$

The parameter α_{disto} depends on the linearity specification and the distortion performance of the circuit or building block and is smaller than or equal to 1. After substitution into (3.111) and (3.61), the quality of the one transistor voltage amplifier becomes:

$$\frac{\mathrm{Gain}^2 \mathrm{BW} \mathrm{Acc}_{\mathrm{rel}}^2}{P} = \alpha_{disto}^2 \cdot \frac{V_{DD}}{24\pi(V_{GS} - V_T)A_{VT0}^2 C_{ox}} \tag{3.113}$$

The more relaxed the linearity specifications are the closer α_{disto} approaches unity and the better the total performance of the design becomes. For high linearity constraints the α_{disto} factor becomes small and as a result the power consumption for a given gain, speed and accuracy performance strongly increases due to the linearity requirements. For current processing stages a similar analysis can be done and the same conclusions can be drawn.

For the performance analysis of building blocks in this chapter, optimistic assumptions have been used in view of distortion performance and linearity specifications for the maximal attainable signal levels. In circuits that must meet high linearity specifications, the impact of distortion will strongly reduce the maximal signal levels and results in a lower dynamic range and a lower accuracy for the same current consumption and supply voltage. Consequently the linearity specification further increases the power consumption of the system to attain a specified speed and accuracy. In this perspective, the theoretical limits derived in this chapter, become even more difficult to approach in practical signal processing systems.

3.8 Implications for analog parallel signal processing systems

Accuracy specifications

The results and analysis presented in this chapter are very important for the design of analog parallel signal processing systems. In their VLSI realization we want to achieve a high density combined with a low power consumption. Both objectives are influenced by the implication of transistor mismatch. The area of the circuits and thus their density is determined by the accuracy specification; but also the power consumption for a given speed is determined by the accuracy specifications since the performance ratio $Speed{\cdot}Accuracy^2/Power$ is fixed by technology constants. A good accuracy specification of analog signal processing is thus very important; an over-specification results in poor speed and power performance and too loose accuracy specifications will result in faulty system operation. Therefore in the next chapter the generation of good accuracy specifications for the building blocks is treated in detail and the necessary theoretical evaluation methods are derived.

Since the area of a circuit is proportional to the $\text{Accuracy}^2 \cdot A_{VT0}^2$, we can rewrite the ratio $Speed{\cdot}Accuracy^2/Power$ as:

$$\frac{\text{Speed}}{\text{DensityPower}} \propto \frac{1}{C_{ox}} \tag{3.114}$$

so that we clearly see that the total performance of an analog parallel signal processing system is also limited by a technology constant of the used VLSI technology.

Weak Inversion Operation

The study of the optimal design of basic building blocks shows that current processing blocks should be biased with large gate-overdrive voltages ($V_{GS} - V_T$) and voltage processing blocks should be biased with small gate-overdrive voltages and if possible in weak-inversion. This analysis thus shows that weak inversion operation of transistors is interesting for the VLSI implementation of parallel systems and neural systems but only for the voltage processing and the voltage mode blocks. Any current processing block must be biased in strong inversion to obtain good performance. An OTA circuit for instance should be designed with its input transistors in weak inversion but the biasing and current mirror stages should use strong inversion as much as possible. A more correct statement is thus that for the implementation of analog parallel systems an optimal combination of weak and strong inversion biasing is to be used.

Analog or digital implementation

In order to decide whether analog or digital circuits are best suited for the implementation of massively parallel signal processing systems we have to compare their speed, power and accuracy performance or equivalently their power/speed ratio as a function of the dynamic range. In figure 3.22 the fundamental limits on the energy per cycle for analog circuits imposed by transistor mismatch and thermal noise are plotted; also the energy per cycle for

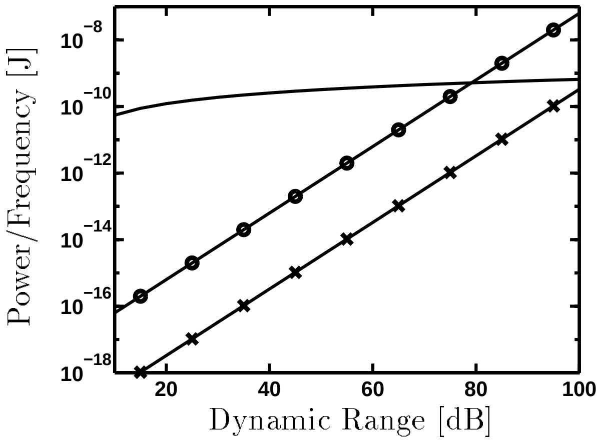


Figure 3.22: The energy consumption for a digital implementation (–) of a pole as a function of the required dynamic range; the minimal energy consumption of analog implementations due to the impact of mismatch (–o–) or noise (–x–).

a typical digital implementation of a single pole transfer function is plotted after [Vit 90b]. The power consumption curve of analog circuits has a steeper slope but starts from a much lower intercept; the curve for the digital circuits is only logarithmically dependent on the dynamic range, since increasing the dynamic range basically only requires the addition of an extra bit, but it starts from a much higher intercept.

Analog circuits are thus clearly advantageous for the implementation of systems with a low dynamic range requirement or low accuracy specifications. This type of specifications typically are required by massively parallel systems which perform perception tasks since they obtain their overall performance from the parallellism rather than from the high quality of the building blocks. Digital circuits are however clearly preferable when high dynamic ranges are required as in typical signal restitution tasks [Vit 90b, Vit 94]. However, transistor mismatch considerably lowers the dynamic range limit below which analog systems perform better compared to the limit imposed by thermal noise.

For the implementation of massively parallel signal processing systems analog circuits offer the best performance. However, the implications of transistor mismatch must be carefully taken into account; therefore good accuracy specifications are extremely important to obtain a high quality chip implementation.

3.9 Conclusions

This chapter discusses the implications of transistor mismatch on the design and on the performance of analog circuits and systems. First a characterization method is developed to extract the parameters for quantitative mismatch models. Since transistor mismatch is caused by statistical phenomena, a large number of sensitive measurements have to be performed using dedicated test circuits and a dedicated automatic measurement setup. They show an inversely linear dependence of the parameter mismatch on the gate area of the transistors. When migrating to sub-micron and deep-sub-micron technologies, the influence of the short and narrow channel effects have to be accounted for in model extensions.

This firm quantitative understanding of transistor mismatch forms the basis for the analysis of its impact on circuit and system performance. We prove that the maximal total performance or the performance ratio *Speed·Accuracy*^{2}*/Power* of elementary voltage and current building blocks is determined by the chosen biasing point – the gate-overdrive voltage $(V_{GS} - V_T)$– and the technology matching quality. A current processing stage must be biased with large $(V_{GS} - V_T)$ to obtain the best total performance whereas voltage stages must be biased with low $(V_{GS} - V_T)$'s as long as the bandwidth requirements allow it. Under optimal biasing conditions, the total performance or *Speed·Accuracy*^{2}*/Power* of the stage is fixed and inversely proportional to the technology mismatching expressed by $C_{ox}A_{VT0}^2$; the lower the $C_{ox}A_{VT0}^2$ of a technology the better the circuit performance. This result thus explicitly states that a circuit designer can only trade one specification for another; but the best total attainable performance i.e. a combination of a high speed and a high accuracy at the same time as a low power consumption, is limited by the implication of transistor mismatch.

These results are then extended for more complex circuits, including opamps and feedback systems. Moreover, for a general analog signal processing system, we show that transistor mismatch again puts a fundamental limitation on the minimal power consumption for a given frequency and accuracy performance. This technological limitation is even several orders of magnitude more important as the physical limitation imposed by the effect of thermal noise. For high speed and massively parallel analog systems and any other analog system where no offset compensation or calibration can be done, mismatch is the performance limiting effect, and the system design must carefully take mismatch into consideration.

The analysis of the scaling of the mismatch behavior with the technology feature size, shows that the evolution towards deep-sub-micron technologies improves the matching quality of the technology. However, the smaller power supply voltages in scaled-down technologies limit the available voltage swing and reduce the scaling advantage. The more pronounced short channel effects and velocity saturation effects also deteriorate the performance of analog circuits and reduce the scaling benefits. Furthermore, for very deep-sub-micron technologies ($< 0.25\ \mu m$), the current factor mismatch prevent a further quality improvement so that further down-scaling will have no positive effect as far as the presently available data indicates.

At the end of the chapter, the implications of these results for the design of massively parallel analog signal processing systems are discussed. Due to the fixed ratio of performances, the importance of good accuracy specifications is demonstrated. The generation of accuracy specifications requires a sound understanding of the impact of random errors on the system level, a subject which is treated in the next chapter. Moreover, we show that analog circuits are indeed more power efficient for low accuracy levels than digital implementations, but mismatch lowers the boundary where digital implementations take over. We also indicate that the sub-threshold operation of transistors, which is generally believed to be the best regime for massively parallel analog systems, is not optimal for all circuit functions.

It is important to stress that the methods and analysis presented in this chapter are generally valid for the design of many types of analog systems. Transistor mismatch is an important factor in the design of analog circuits and therefore a good quantitative transistor mismatch model for the used technology should always be available to the analog designer. A good knowledge of mismatch allows a better optimization of the circuit design and avoids that very large safety margins have to be taken in the design which result in poor power and speed performance.