POLITECNICO DI TORINO

Department of Electronics and Telecommunications
Master's degree in Electronic Engineering

Thesis

Analysis and Design of a Low Power Analog-to-Information Converter

Author
Carmine Paolino

Supervisor
Prof. Gianluca Setti

Co-supervisor
Prof. Fabio Pareschi

A.Y. 2018/2019

Abstract

The recent growth of the personal medical devices and the precision medicine markets has renovated the interest in searching extremely low power signal acquisition solutions. Biomedical applications, in fact, require battery-operated devices which, especially if implanted in the patient, do not easily lend themselves to being recharged or replaced.

In the system-level design field, it is known that the performance gains achieved by a certain solution are the largest the more abstract is the view on the system. That is why the signal processing research community has spared no effort in finding alternative sampling techniques which work at sub-Nyquist rates and still guarantee the correct reconstruction of the original signal. Hardware architectures with these features are called Analog-to-Information (A/I) Converters and are typically based on the theory of Compressed Sensing (CS).

CS theory describes how a particular families of signals can be recovered from a limited set of linear measurements, using a smaller number of samples than what would be required by a Nyquist-rate approach. It potentially allows a signal acquisition rate much lower than the Nyquist one, saving considerable energy and bandwidth. This is indeed the defining feature of an Analog-to-Information converter.

In this thesis, a new A/I architecture is proposed. It is derived as an extension of traditional charge-redistribution Successive Approximation Register (SAR) A/D converters. By avoiding the introduction of additional power hungry, active elements, it achieves a significant increase in efficiency with respect to other CS architectures proposed in the literature.

The study of device non-idealities has lead to a robust topology, whose design considers as early as possible the actual limitations of the available technology. Several models have been devised on the different circuital blocks, trying to make them as generally applicable as possible. Algorithmic and circuital simulations have shown the ability of the architecture to closely match ideal reconstruction performances, overcoming the major drawbacks identified during the initial analysis phase, and achieving good performance figures.

Introduction

Signal acquisition is traditionally associated to the Shannon sampling theorem, which defines a lower bound to the frequency at which signal samples are collected. Although it is a general theory that only depends on the frequency spectrum of the signal of interest, it results in a non-optimal bound for several signal families, requiring excessing resources and resulting in unnecessarily high power consumption.

The theory of Compressed Sensing (CS) is an active research field in signal processing which enables simultaneous sampling and compression of a broad family of signals. Architectures employing this framework, thus working at a rate below the one defined in Shannon's theory, are called Analog-to-Information converters.

Since the work by Candès, Romberg and Tao, the engineering community has spared no effort in the design of effective CS signal processing chains. Many solutions have been proposed on CS theoretical aspects to make more efficient both the acquisition and reconstruction more efficient, and results on CS-based sensing circuits/systems have emerged, too.

Identifying discrete-time signals as vectors in $x \in \mathbb{R}^n$, CS describes how the original information vector x can be recovered from a limited set of linear measurements $y = Ax \in R^m$, where the sensing matrix $A \in R^{m \times n}$ and $m < n$. The requirement is for the signal to be sparse in a given basis, that is, to have only a few significantly non-null coefficients in its representation. The number of measurements, smaller than what required by a Nyquist-rate approach, potentially allows a signal acquisition rate much lower than the Nyquist one, saving considerable energy and bandwidth.

In this thesis, a novel architecture is proposed. It relies on the capacitive array found in traditional Successive Approximation Register (SAR) A/D converters. The advantage is the possibility to perform all the operations of CS-based sensing (i.e., a single row-by-column product of Ax and the digital conversion of the result) using the active elements already present in the converter, with a significant increase in efficiency with respect to other CS architectures proposed in the literature.

The underlying idea is to consider antipodal sensing matrices $A \in \{-1, +1\}^{m \times n}$ only, thus reducing the multiplications involved in the product Ax to trivial sign inversions. Either x or $-x$ is then sampled at different time instants on the capacitors of the SAR array, and the sum is obtained by charge redistribution. The main issue is the limited hold-time capabilities of the (small) capacitive cells, due to the leakage currents in the pass transistors. Two approaches have been investigated to alleviate the problem: modifying the sensing matrix A, making it block diagonal, and compensating the information loss through a hardware feedback loop.

Using a block diagonal sensing matrix, each row-by-column product involves only a fraction of all the samples in a signal window. From a circuital point of view, the advantage is twofold. Complexity of the array is reduced and the shorter acquisition window relaxes the hold-time requirements. Since this might impair the reconstruction quality, extensive software simulations have been performed to characterize the effect, and to establish the proper range of block sizes that would result in good reconstruction quality. Additionally, a low-power leakage compensation circuit, requiring a feedback loop around every hold capacitor, has been introduced and studied both in the time-domain and in terms of stability.

The thesis is divided in two parts. Part I is concerned with a system level description of the mathematical theory underlying this work, as well as the description of the proposed solution. Part II is entirely focused on the investigation of device non idealities from a circuital point of view, leading to the definition of several design constraints.

In Chapter 1, the theory of sparse signal recovery and, by extension, Compressed Sensing, is introduced. They are then applied to electrocardiogram (ECG) signals, highlighting immediately the main limitations such signals pose on the acquisition chain. The proposed A/I architecture is introduced in Chapter 2.

Part II starts with the analysis of the capacitive array employed in the proposed architecture in Chapter 3, in particular with regards to the effects of parasitics. Chapter 4 is a brief review of the typical implementation of pass transistors in a switched-capacitor circuit. In Chapter 5 the leakage compensator is introduced. Its time-domain description, as well as the stability limits have been derived and its effect on the quality of the reconstructed ECG signals is evaluated. Chapter 6 is concerned with the description of the comparator and the derivation of the input-referred offset voltage. Finally, in Chapter 7, the effect of layout strategies on the mismatch of identical integrated devices is investigated through the use of a simplified statistical mismatch model.

Acknowledgements

Several people where involved, in a way or another, in the technical and personal effort that this thesis has been. Though naming all of them would be impossible, the contribution of some has to be recognized as being paramount in the achievement of a (hopefully) successful outcome.

First of all the group of Prof. Setti, with Profs. Pareschi and Mangia on the first line of attack against all possible setbacks and always present to direct the research effort and change the course of action when appropriate. Then my colleagues and friends (Andrea and Aldo, you belong here!), whose fruitful discussion have lead to interesting implications. And finally my dear brother, always against my intuitive grasping of concepts and lack of mathematical rigor, but always there when I needed to clarify my mind when some subtle difficulty arised.

To all of you, thanks for being there in the moments of need and for stimulating me to

achieve higher goals!

Part I.

Analog-to-Information Conversion

1. Compressed Sensing

Compressed Sensing (CS) is a signal processing technique able to reconstruct particular signal families using far fewer measurements than what the Shannon sampling theorem suggests. This chapter will delve into the mathematical theory of CS, explaining the reasons of its effectiveness and discussing some theoretical bounds defining the limits of applicability.

Rakeness-based CS is then introduced as a way to improve the performance of the technique, evaluated in the case of electrocardiogram (ECG) signals. Finally, the use of hardware-friendly sensing matrices will be analyzed in order to simplify the design of a circuital implementation of a so called Analog to Information converter.

1.1. Why bother?

Whenever the need to acquire a continuous-time signal arises, Shannon sampling theorem enters the discussion [1]. The theorem states that functions whose Fourier spectrum is null above a certain frequency f_{max}, also known as the bandwidth or the Nyquist frequency, can be described exactly from samples collected at least at twice that frequency. As the operating speed of information-processing systems is increasingly higher, complying with this bound becomes unfeasible because of technological limitations. At the same time, the frequency domain is not necessarily the one that shows the most striking features of a signal, since the true information rate might not be readily observable (a fixed-frequency sine wave, does not provide new information over time, even if it is varying continuously), leading to an unnecessarily high Nyquist bound. Weakening such a constraint would lead to dramatic reduction of the resources employed.

The main advantage of the Nyquist-rate approach is that it is valid under general conditions, i.e. for any signal having limited bandwidth. Therefore, once a system has been designed to operate at a given sampling frequency, more often than not, it can manage any signal whose Nyquist frequency is low enough. Furthermore, if the sampling frequency f_s is sufficiently larger than the Nyquist one (as a rule of thumb $f_s > 10 \times f_{max}$), a reasonable approximation of the original information can be recovered by a simple linear interpolation.

In practice, any signal processing chain will be designed for a specific application. The operating conditions, like the kind of signals involved, will therefore be known in advance. Since some parameters already have to be matched to the specific properties of the signals (e.g. the sampling frequency), one might ask if it couldn't be possible

to further tune the system properties to take advantage of more prior information. As an example, consider a pure sine wave at frequency f, Shannon theorem requires a sampling frequency greater than $2f$ to be able to recover the original information. However, having prior knowledge of the fact that the signal is indeed a sinusoid, only three measurements in total are sufficient to define its amplitude, frequency and initial phase. Therefore including in the design of a signal processing chain information on the properties of the signal family, the acquisition effort can be significantly reduced. What Compressed Sensing achieves is to observe the signal in a different domain and, using a property of the signal family called sparsity, work with far less scalars than what the Nyquist constraint would require.

Compressed Sensing has been developed as an extension of the theories on the recovery of sparse signals, with the distinction that CS processes the samples in a domain where the signal is not sparse. Therefore we will first analyze the properties and bounds involved in sparse signal recovery, later moving to the features of CS.

1.2. Sparse Signals Recovery

While we typically work on continuous-time signals, Compressed Sensing theory is most simply understood in a discrete, finite-dimensional setting. In that context, signals can be thought of as sequences of n Nyquist-rate samples, i.e. vectors in $\mathbb{R}^n$. A basis of such a space is the smallest set of vectors able to represent any other element by a proper linear combination. Although the number of basis for a given space is unlimited, each signal has a unique representation in any of them. If the matrix $\Phi \in \mathbb{R}^{n \times n}$ contains, on its columns, the basis vectors, a generic element x of the space can be expressed as

$$x = \Phi \xi. \tag{1.1}$$

The vector $\xi \in \mathbb{R}^n$ is an equivalent way to look at the original vector. In geometrical terms, the product in (1.1) is a change of coordinates. It preserves the informative content, with the possibility to observe other features of the original vector.

The fundamental property required by the theory underlying Compressed Sensing is *sparsity* of the signal. A vector in $\mathbb{R}^n$ is κ-sparse if the number of non-null coefficients is $\kappa \ll n$. In general, the linear combination of two κ-sparse vectors is at most 2κ-sparse, since the non-null coefficients may occupy different positions. Knowing the sparsity level of a vector, the quest is to reduce the number m of observations required to describe it uniquely such that $\kappa < m \ll n$, providing sufficient information to recover the original data.

Consider $\xi \in \mathbb{R}^n$ as a κ-sparse sequence of length n. A *measurement* corresponds to the linear combination of the n scalars in ξ, weighted by some a_i coefficients, and resulting in a single number

$$y_j = \sum_{i=1}^{n} a_i \xi_i \qquad j = 1, \dots, m.$$

In Nyquist-rate sampling, a measurement would correspond to a single sample, while in this context it involves the processing of an entire window of length n. The evaluation of m such measurements can be compactly written as

$$y = A\xi, \tag{1.2}$$

where the a_i coefficients are placed on the m rows of $A \in \mathbb{R}^{m \times n}$. The mapping described by A is from the n-dimensional space to a lower dimensional one.

Recovering ξ from y is, in general, not possible since the dimensionality reduction implies that multiple ξ map to a single y. Equivalently, this process corresponds to solving an underdetermined system of equations which, according to Rouché-Capelli's theorem [2], has ∞^{n-p} solutions, with p the rank of the matrix (the number of linearly independent columns) and $n - p$ being the number of free variables. Since $p \leq m$, the number of free variables is greater than or equal to $n - m$. In reality, by having prior knowledge on the sparsity of ξ, uniqueness of the solution can in fact be guaranteed.

1.2.1. Uniqueness conditions

Intuitively, knowing that the ξ's of interest are κ-sparse, any of them should be distinguishable from the others after the projection into the lower dimensional space. That is, for any ξ_1, ξ_2, the difference $\Delta\xi = \xi_1 - \xi_2$ observed in the target space has to be

$$\Delta y = A\Delta\xi \neq 0.$$

The term $\Delta\xi$ is, in general, 2κ-sparse, and the non-null positions are unknown. Considering Δy as the linear combination of columns of A, weighted by the coefficients in $\Delta\xi$, we have to guarantee that any 2κ columns of A are linearly independent. This way, there is no possibility of obtaining the null vector from the weighted sum of the matrix columns and, as a result, any ξ is still distinguishable in the smaller, m-dimensional space. Identifying uniquely the original ξ, starting from y, is then possible, with the advantage of having to perform only $m \ll n$ observations.

This qualitative description can be formalized by introducing the concept of *spark* of a matrix [3]. It is the maximum cardinality c such that any subset of c columns of A contains only linearly independent elements Therefore, recovering uniquely a κ-sparse vector $\xi \in \mathbb{R}^n$ from m linear observations ($m < n$) is possible if

$$\kappa = \|\xi\|_0 < \frac{1}{2}\text{spark}(A), \tag{1.3}$$

where $\|\cdot\|_0$ is the ℓ_0 (pseudo)norm, equal to the number of non null coefficients in the vector.

Under this uniqueness condition, the solution of (1.2) can be found through a minimization process that looks for the sparsest ξ such that $y = A\xi$, i.e.:

$$\begin{aligned} &\underset{\xi \in \mathbb{R}^n}{\text{argmin}} \|\xi\|_0 \\ &s.t. \quad y = A\xi. \end{aligned} \tag{1.4}$$

However, computing the spark requires a combinatorial search (NP-hard) over all possible subsets to evaluate the independence of their elements. A more easily computable index resulting in a uniqueness condition for the solution of (1.2) is based on the concept of *mutual coherence* of the columns of A, defined as:

$$\mu(A) \stackrel{\text{def}}{=} \max_{i<j} \frac{\left|A_i^T A_j\right|}{\|A_i\|\|A_j\|},$$

where A_i represents the i-th column of matrix A. Mutual coherence is constrained in the range $0 \leq \mu(A) \leq 1$. It is equivalent to the maximum cosine of the angle between any two columns, corresponding to the smallest acute angle between them. A high coherence is therefore equivalent to aligned vectors. From a different point of view, considering the columns as random vectors, mutual coherence represents the maximum correlation coefficient between the columns.

Since the mapping represented by A should store as much information as possible with the lowest redundancy, we expect a low coherence to result in better performance. Equivalently, the columns should be as orthogonal as possible or, looking at them as random vectors, as uncorrelated as possible. Indeed the uniqueness condition in (1.3) can be stated in terms of the mutual coherence index, becoming

$$\|\xi\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right).$$

This theoretical bound, representing the range of sparsity for which a unique solution can be found, is effectively increased by a lower coherence of A. However, since it can be shown [3] that

$$\text{spark}(A) \geq 1 + \frac{1}{\mu(A)},$$

the new bound is lower than the one based on the spark, thus it is applicable on a smaller subset of vectors, which have to be even sparsest.

1.2.2. Numerically tractable recovery

Up to now we have used the sparsity level of the vectors, quantified by their ℓ_0 norm, to find the correct solution to the minimization problem. Such a computation requires, however, a combinatorial search across all possible candidates and is not suitable for an efficient numerical implementation. An important consequence implied by the mutual coherence number is the equivalence of the result found by using the ℓ_1 norm instead [3]. The ℓ_1 norm corresponds to the summation of the absolute value of all vector coefficients and it is the smallest-rank norm that is convex, therefore suitable for use in a numerical

implementation. The reconstruction in (1.4) then becomes:

$$\begin{aligned} &\underset{\xi \in \mathbb{R}^n}{\operatorname{argmin}} \|\xi\|_1 \\ &s.t. \quad y = A\xi, \end{aligned} \tag{1.5}$$

where the only change is the rank of the norm. In turn, this forces stricter conditions on the sensing matrix A, which has to be generated with greater care.

Although this formulation leads to a (possibly) efficient numerical implementation, empirical evidence shows that the sparsity range for which the unique solution can be found is actually larger than what the mutual coherence implies. New indexes have been proposed, the most popular one being the *isometry constant*, defined as the smallest number such that, for all κ-sparse vectors ξ, the following property holds:

$$(1 - \delta_\kappa)\|\xi\|_2^2 \leq \|A\xi\|_2^2 \leq (1 + \delta_\kappa)\|\xi\|_2^2.$$

This condition is also referred to as the *Restricted Isometry Property* (RIP). It implies that the projection into the m-dimensional space approximately preserves the norm of the original vector.

Similarly to the reasoning on the meaning of the spark, the goal is to distinguish any sparse ξ, therefore the distance $\Delta\xi$ should be preserved and the isometry constant of interest is actually $\delta_{2\kappa}$. It can be shown that for $\delta_{2\kappa} < 1$, the ℓ_0 minimization can find the unique κ-sparse solution. The bound becomes $\delta_{2\kappa} < \sqrt{2} - 1$ in the case of the ℓ_1 minimization, highlighting the fact that using a more convenient norm in the optimization process requires a more careful design of the measurement operator A.

The three indexes introduced thus far can be linked. Indeed if $\delta_{2\kappa} < 1$, any subset of 2κ columns of A has linearly independent elements, therefore $\operatorname{spark}(A) > 2\kappa$. Moreover, the RIP of order κ is satisfied, with $\delta_\kappa \leq (\kappa - 1)\mu(A)$.

1.2.3. Sensing matrix design

The construction of the sensing operator A in order to achieve a small coherence or a small isometry constant is of great concern. This would ensure a wide range of sparsity in which a unique solution can be found. A deterministic process leading to matrices with suitable properties would involve the solution of

$$\underset{A \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \mu(A) \qquad \text{or} \qquad \underset{A \in \mathbb{R}^{m \times n}}{\operatorname{argmin}} \delta_{2\kappa}(A).$$

Both these problems do not lend themselves to an easy evaluation, being of combinatorial nature. Indeed it can be shown that these properties hold with probability close to 1 for random matrices of size $m \times n$ [4]. In such matrices, each entry is a realization of some random variable, with a chosen probability density function (e.g. Gaussian,

Uniform, Bernoulli). The simplest case is for the entries to be independent and identically distributed (i.i.d.). Moreover, the number of measurements m cannot be too low. Indeed

$$m = \mathcal{O}\left(\kappa \ln \frac{n}{\kappa}\right),$$

where the proportionality depends on δ_κ. Still, since the lower bound on m could be higher than what is actually required for correct reconstruction, the empirical study of a reasonable range for m might be a suitable way to approach the problem at the beginning.

An alternative way to look at the use of random variables is that this ensures a high degree of spreading of the matrix columns in the signal space, with the further advantage of being robust against the loss of part of the measurements.

1.2.4. Noisy measurements

All the previous theoretical guarantees have been obtained in the context of noiseless measurements and exact κ-sparsity of the original vector. If any of these two conditions is not met, we need to ensure that the tools developed so far can still provide the expected results. Indeed, if the measurements vector is affected by a noise term η

$$y = A\xi + \eta,$$

the solution to the minimization problem is no more exact and (1.5) becomes

$$\begin{aligned} &\min_{\xi \in \mathbb{R}^n} \|\xi\|_1 \\ \text{s.t.} \quad &\|y - A\xi\|_2^2 \leq \varepsilon \end{aligned}$$

The solution $\hat{\xi}$ has to result in a projection $\hat{y}$ close to the measured one y, with its uncertainty depending on the amount of noise $\varepsilon(\eta)$. The same is true if the original signal is not sparse but *compressible*, in which case most of its coefficients are small but non-zero. Reconstruction based on sparsity won't be able to result in the same exact measurement acquired from the signal, but will be close enough.

Moreover, the uniqueness of the solution and the equivalence of the ℓ_0 and ℓ_1 minimization procedures no longer apply, but it can be shown that the RIP condition guarantees robustness of the formulation against noise [3].

1.3. Enter Compressed Sensing

In the previous discussion, measurements were computed directly from sparse vectors. In reality, sparsity is not necessarily observed in time domain, therefore if one wants to apply the previous results, the observation have to be performed on $x = \Phi\xi$, where

$x \in \mathbb{R}^n$ is a signal window containing n Nyquist-rate samples, Φ is the $n \times n$ sparsity basis whose columns are the coordinate vectors of the space and ξ is κ-sparse. The measurements vector y is obtained by applying a linear projection $S \in \mathbb{R}^{m \times n}$ to the vector x (containing the time-domain samples), such that

$$y = Sx = S\Phi\xi.$$

It is the composition $S\Phi$ that has to satisfy the RIP of order 2κ. If the number of observations m is sufficient, then a unique solution ξ can be found by the solving

$$\begin{aligned} &\underset{\xi \in \mathbb{R}^n}{\operatorname{argmin}} \|\xi\|_1 \\ \text{s.t.} \quad &\|y - S\Phi\xi\|_2^2 \leq \varepsilon \end{aligned}$$

Then x can be recovered knowing Φ. The advantage introduced by Compressed Sensing is the early reduction of the number of measurements to be acquired, thus avoiding the need to collect every single sample. The *Compression Ratio* quantifies this gain, being expressed as

$$\mathrm{CR} \stackrel{\text{def}}{=} \frac{m}{n}.$$

Since the conditions on A are actually posed on the product $S\Phi$, but the matrix to be designed is S, it is necessary to look into how the properties of S are carried over to the compound operator. If Φ is orthonormal, then building S according to a Gaussian distribution and satisfying the RIP will result in a product $S\Phi$ with the same properties [3]. The same is true if the columns of S have low coherence to the columns of Φ. These consideration can also be extended to sub-gaussian distributions [4].

1.3.1. Rakeness-based Compressed Sensing

Continuing along the path of specialization of the signal processing chain, a significant improvement in reconstruction performance can be achieved by observing another feature of the signal family. Intuitively, once the sparsity basis has been identified, not necessarily the coefficients associated to each basis vector have the same average length, i.e. energy. Equivalently, the signal instances in the n-dimensional space are not uniformly spread, but concentrate along some of the directions. By focusing the measurements on the more energetic directions, the average energy collected (*raked*) by each measurement can be maximized, with a significant gain in the quality of the reconstructed signal.

The quantity measuring the distribution of energy across the basis vectors is named *localization* [5] and it represents, together with the sparsity level, an additional prior to the reconstruction process. Its effect is observed on the correlation profile of the sensing matrix rows:

$$C_S = \frac{1}{2}\left(\frac{C_x}{\mathrm{tr}(C_x)} - \frac{I_n}{n}\right),$$

where I_n is the $n \times n$ identity matrix, $\mathrm{tr}(\cdot)$ the trace operator and $C_x = \mathbb{E}[xx^{\mathrm{T}}]$ the expected correlation profile of the signal. C_x can be evaluated either having a model of the signal (and running some Monte Carlo simulations) or having acquired a large dataset of signal waveforms.

This simple modification to the generation of the sensing matrix leads to dramatic improvements in the reconstruction quality, therefore becoming essential to minimize the necessary resources.

1.4. Sensing of ECG signals

The application of interest for this work is the acquisition of ECG signals. Applying CS techniques to this kind of signals requires the determination of the right sparsity basis, the optimal window length so that a sufficient level of sparsity is actually observed and finally the spreading of energy across the basis vectors (localization). Such properties cannot be evaluated analytically.

The most important feature we are looking for in a candidate basis is the sparse representation of the signal. According to results published in [6], the Symmlet-6 basis is a suitable candidate. Fig. 1.1 shows the shape of the mother wavelet and the two lowest order siblings for such a basis [7]. It is clear why ECG-like signals can be sparsely represented in terms of these functions because of their mutual resemblance.

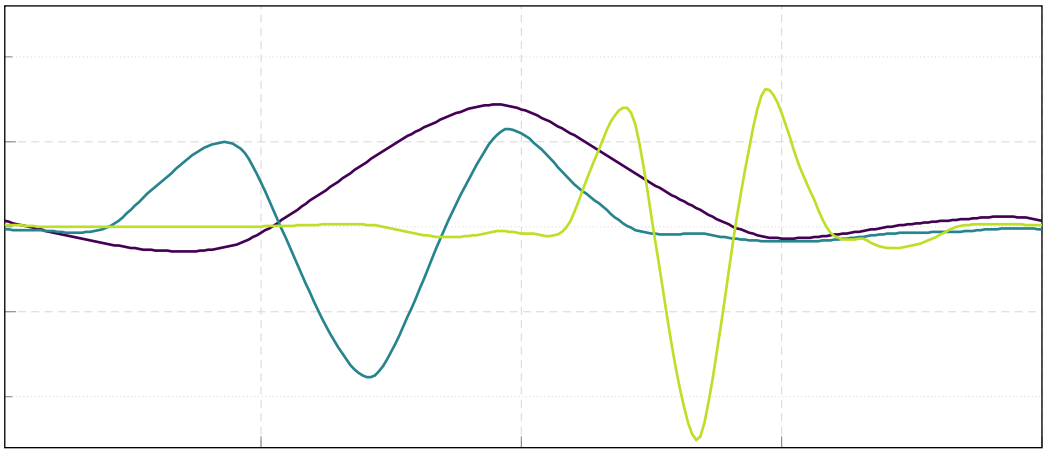

Figure 1.1.: Symmlet-6 basis functions. Mother wavelet (darkest) and first two siblings.

By considering each window of the signal as the realization of a random process, it is clear that sparsity will be a random variable with some probability distribution. By choosing a proper size for the signal window, the variance of such a variable can be reduced, obtaining consistent quality on any window. Intuitively, if the window to be processed is too small (less than a period of ECG), than the signature features of an

ECG waveform are lost and the representation of some sequences (the most relevant ones) might not be sparse anymore. A size of at least one ECG period is thus required. Since sparsity is dependent on the time duration of the window (at least one period), the number n of values processed by the CS acquisition has to be related to the sampling frequency, i.e. the number of signal samples in a window of a specified duration. The results in [8] show that $n = 256$ is a good compromise between complexity of the reconstruction (which scales as n^2), if the sampling frequency is at least 256 Hz.

The instances of ECG signals used in this work have been generated according to the model defined in [9]. In Fig. 1.2 is depicted one ECG period against the reconstructed waveforms obtained using standard and rakeness-based CS, with $n = 256$ and increasing compression ratios.

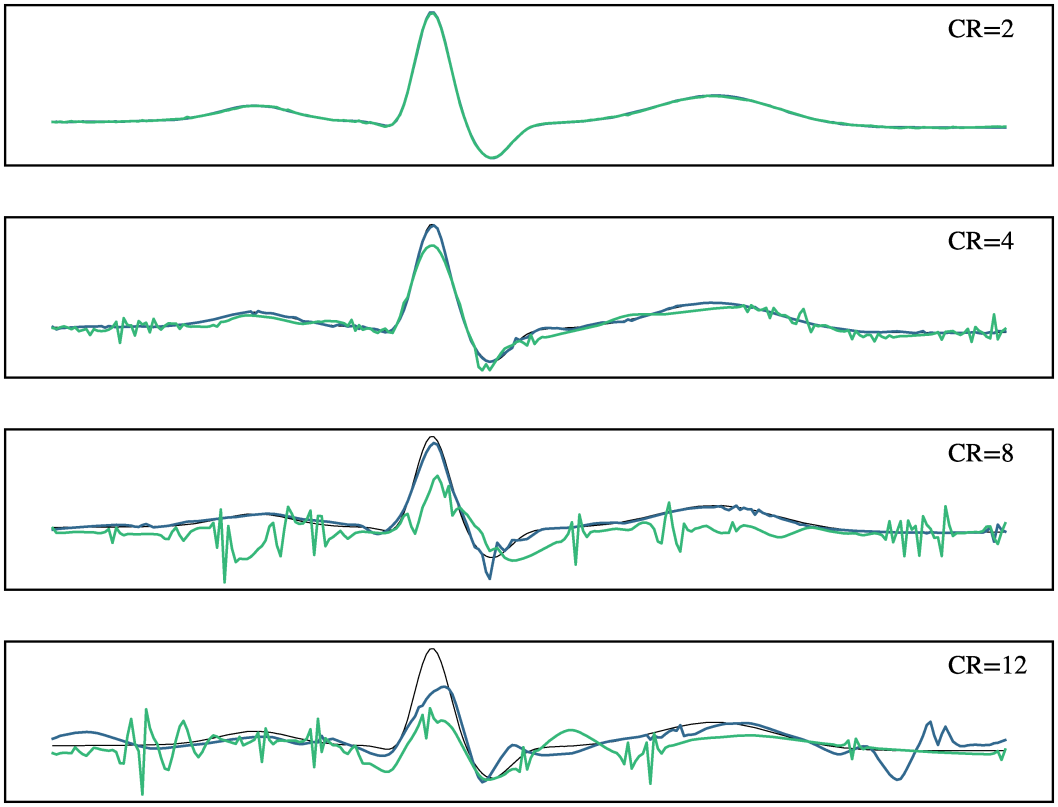

Figure 1.2.: Reconstructed ECG for different compression ratios, against the original waveform. Comparison between standard CS (green) and rakeness-based CS (blue).

As CR is increased, the reconstructed waveforms are affected by an increasingly higher level of noise. The ones obtained using rakeness-based CS still allow a clear distinction of the main peak even at high compressions. In applications where it is only necessary

to evaluate the cardiac frequency, this result might be sufficient.

1.5. Designing Efficient Sensing Matrices

An efficient sensing matrix is one that leads to a simple hardware realization of the acquisition process. Since CS is more advantageous the closest it is applied to the signal source, this means having to work in the analog domain. Therefore, having floating-point matrix entries would require, apart from a way to generate said values, the introduction of analog multipliers in the acquisition system, both power hungry and limited bandwidth.

The use of antipodal and ternary matrices, instead would only require sign inversion of the signal, being the values of $A_{i,j}$ constrained to the set $\{-1, 0, 1\}$. We refer to [5] for ways to generate antipodal sequences with a prescribed correlation profile. Here we will only observe numerical results that take advantage of said simplification.

Other than limiting the set of coefficients to be used, a further improvement would be to modify the structure of the sensing matrix in order to reuse the hardware resources. Using block diagonal matrices (Fig. 1.3), each block would process independent parts of a signal window. Therefore the same hardware could be shared among the blocks, being them orthogonal in time. If the blocks have size $m_b \times n_b$, the same resources are used n/n_b times. At the end of each block, the measurements are evaluated and the hardware can start to process the following block of coefficients.

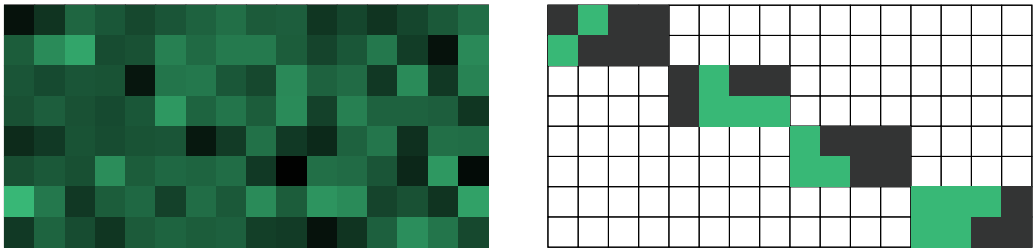

Figure 1.3.: Full 8×16 random matrix (left) vs block diagonal antipodal matrix (right).

As the blocks are made smaller, the number of zeroes in the matrix increases, complexity of the hardware is reduced, though at the cost of quality as shown in Fig. 1.4, where the Reconstruction Signal-to-Noise Ratio (RSNR) quantifies the performance of the CS reconstruction. It is defined as:

$$\mathrm{RSNR}[\mathrm{dB}] = 20 \log_{10} \left(\frac{\|x\|_2}{\|\hat{x} - x\|_2} \right)$$

This parameter compares the energy of the original signal to that of the error with respect to the reconstructed one. The plot shows the Average value of the RSNR (ARSNR)

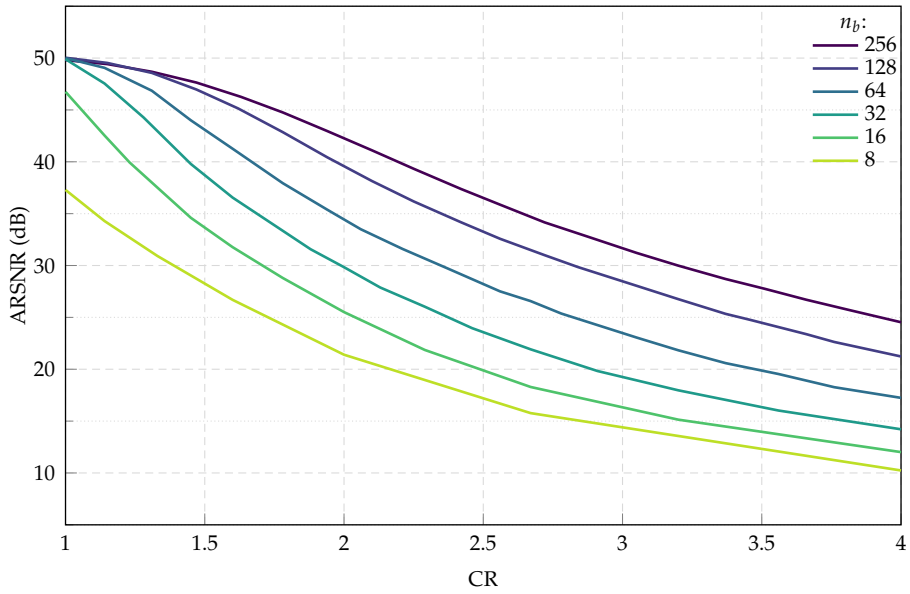

Figure 1.4.: Performance of a CS-based signal processing chain in terms of ARSNR as a function of CR. Synthetic ECG signal at approx 60 beats/s, sampled at 256 Hz and using $n = 256$.

observed over 10000 Monte Carlo trials, where additive noise has been introduced to model non-idealities, starting from an intrinsic SNR of 50 dB. As expected, a higher n_b results in a better quality of reconstruction, with the full matrix providing the best possible outcome.

The choice of n_b involves considerations of several parameters involved in the hardware design. The choice of a proper value will therefore be delayed until the architecture will be defined and some other phenomena observed.

2. Charge Redistribution A/I Architecture

As shown in the previous chapter, Compressed Sensing specializes the sampling operation, matching it to the properties of the signals to be acquired. Here we will propose a low-power analog implementation of the sensing process. It is based on a successive approximation register (SAR) A/D architecture because of its already low energy consumption and the relative ease in adapting it to the requirements dictated by CS. The goal is to perform the matrix-vector product $y = Sx$ in the analog domain, maximizing the use of the passive structures already found in the A/D converter.

First, we will describe the behaviour of the SAR algorithm, followed by the analysis of a switched-capacitors implementation. The modifications applied to the traditional structure are then discussed, highlighting the need to take care of several nonidealities.

2.1. Where to Apply Compression

Compressed Sensing condenses all the meaningful information carried by a signal in as few measurements as possible. However, it can be applied in different points along the signal processing chain, each resulting in its own hardware requirements. Fig. 2.1 shows the main blocks of a typical A/D signal processing chain, highlighting the interfaces at which compression can be performed in order to obtain the desired Analog-to-Information conversion.

The first possibility (Case A) is to work in the continuous-time analog domain. The sensing operation is described by an operator $\mathcal{A}_j\{x\}$ performing a transformation on the input. The most general case involves a modulating function continuous in amplitude, thus requiring some kind of analog multiplier. However working in continuous time, active circuits are required to realize $\mathcal{A}_j$, leading to a significant increase in power consumption. This "CS-first" signal processing chain becomes interesting at extremely high frequencies, where even sampling the input at Nyquist-rate becomes difficult. In that case the compression introduced by CS directly translates to a reduction of the switches' operating frequency.

If instead the input is sampled first (Case B), compression can still be applied in the analog domain. Having discrete-time samples, the transformation becomes a modulation of the input x by some coefficient vectors A_j. With the same considerations as in Section 1.5, the modulating coefficients can be limited to being ± 1, realizing the product as a sign inversion. The sum of the partial terms $A_{j,k}x_k$ involved in the matrix-vector product can be implemented, other than using a discrete time integrator, with an entirely

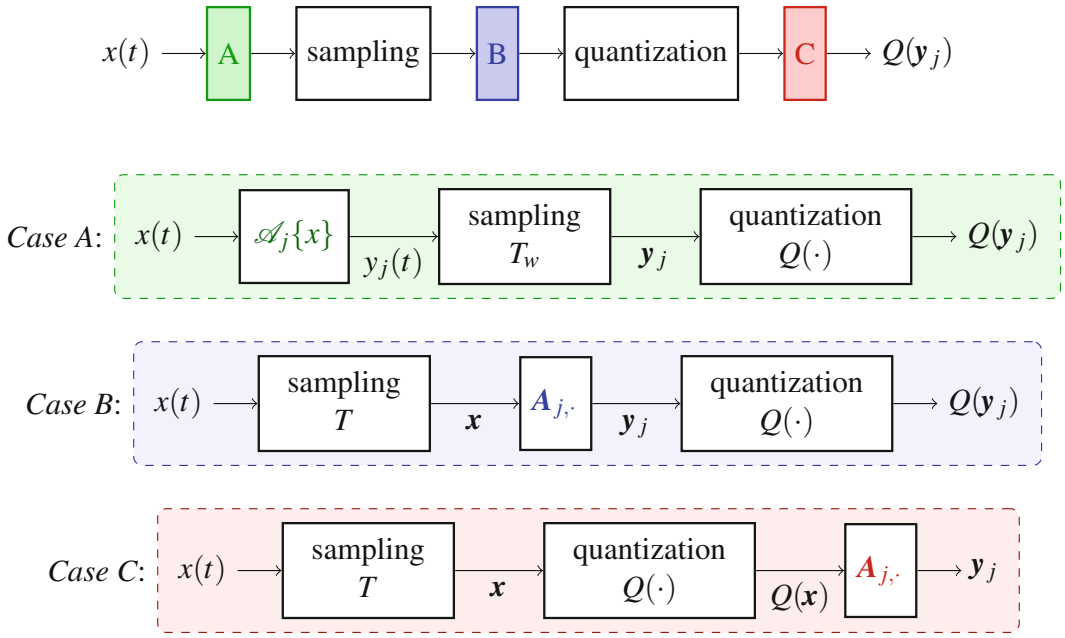

Figure 2.1.: CS along the signal processing chain (reported with permission from [5])

passive solution, as shown in this work. The input switches still work at Nyquist-rate, but A/D conversions are performed only at the end of an acquisition window and extremely low power consumption can be achieved.

The final solution (Case C) involves a digital compression, with no modifications of the ADC block, which generates digital values as usual.

We will focus our discussion on the second solution. Since the core of the architecture is its A/D converter, we will start by describing the conversion algorithm of choice and the traditional topology typically employed in that context.

2.2. Successive Approximation Conversion

The successive approximation algorithm allows the conversion of an analog value in digital form by performing a sequence of comparisons of the input against an adaptive reference [10]. The reference is updated across several cycles, until the required accuracy is reached. If the steps' height decreases as a power of two, a resolution of N bits is achieved in N cycles. A high-level view of the circuit blocks involved in such a conversion is depicted in Fig. 2.2a, with the time-domain behaviour of the most important signals in Fig. 2.2b.

The input signal is sampled so that a constant value is used during the entire conversion. The signal sample is compared against a time-variable reference voltage whose

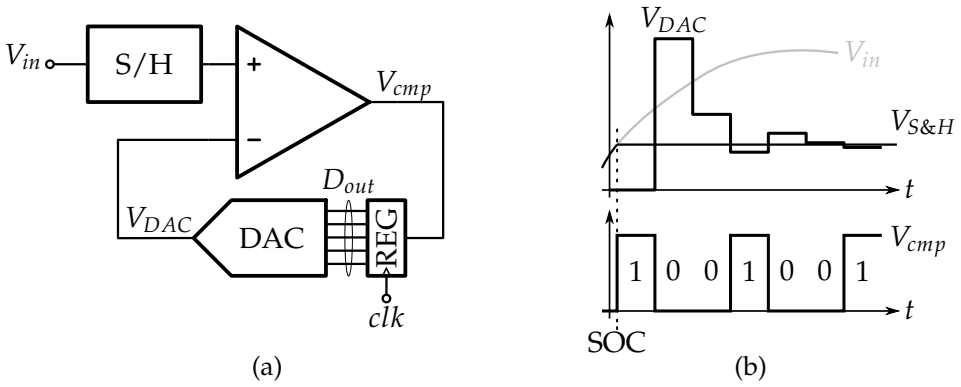

Figure 2.2.: a) Elements of a SAR A/D architecture. b) Signal waveforms during conversion.

evolution depends on the sign of the previous comparison. If the reference voltage was lower than the input sample, the reference is increased and viceversa.

The outcome of the comparison is stored in a digital register, starting from the Most Significant Bit (MSB) at cycle 1 and moving towards the least significant ones as conversion goes on. The register content is initially reset and represents, over time, the always-improving approximation of the input signal, hence the name of Successive Approximation Register (SAR) converter. The analog equivalent of the register content becomes, through a D/A conversion, the comparator reference. Fig. 2.2b clearly shows how the difference between the DAC voltage and the signal sample progressively decreases, beginning from the assertion of the Start-of-Conversion (SOC) signal. It is also possible, as it happens in the figure, that the final error is not the minimum one (which in the example is achieved at the cycle before the last), though it is guaranteed to be below the quantization error by the end of the conversion.

The most critical element required by the A/D converter is actually the D/A converter embedded in it. A popular solution in CMOS technology is to employ switched capacitors, both because of the good technological properties of switches and capacitors in CMOS processes and for the absence of static currents.

2.2.1. Scaled Capacitive DAC

In a low-power CMOS-based implementation, the D/A element is typically based on a scaled capacitive array (Fig. 2.3). Apart from the advantage of being completely passive, it allows the embedding of the sample and hold component with the adaptive reference generation, both required by the SAR algorithm.

In its simplest form, the array is built as a ladder of capacitors, scaled according to powers of two and sharing the top node. The voltage at this node is observed by the comparator to generate the decision in the SAR algorithm. During the conversion each capacitive element is driven, sequentially, to either the positive or negative voltage

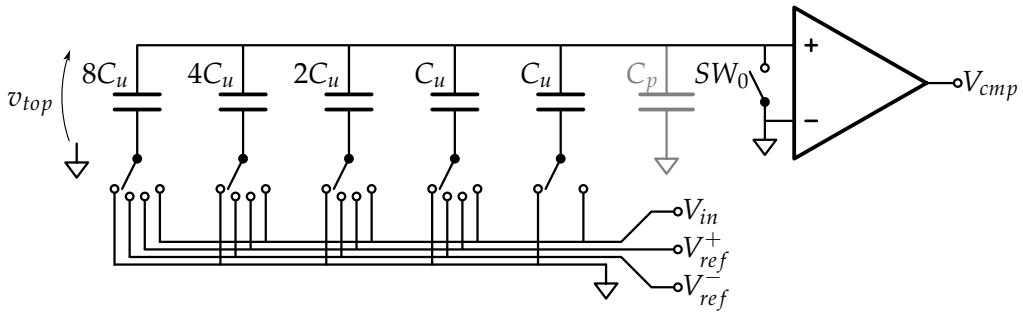

Figure 2.3.: Embedded sampler and D/A converter using a scaled capacitive array. Parasitics are greyed out.

reference, until all bits have been generated.

The sequence of operations starts with sampling. Two alternatives exist on how to manage the array in this phase, namely top- and bottom-plate sampling. Only the latter is applicable in the proposed solution because of the modifications required to implement a CS-based acquisition. The most significant difference between the two techniques concerns parasitics. In bottom-plate sampling they introduce a constant attenuation that can be easily compensated for and does not affect the linearity of the conversion.

The input signal is sampled in the array by holding the top plates at ground (SW_0 closed) and driving simultaneously all the switches on the bottom plates to the input (selectors in Fig. 2.3 on the rightmost position). The capacitance seen from the input is that of the entire array (Fig. 2.4a). SW_0 is then opened, isolating the top plates (Fig. 2.4b) and all the capacitances are grounded. This stored charge makes the D/A conversion signal dependent, as required by the SAR algorithm, but using a single structure. Therefore only one comparator input is occupied and the other can be set to a constant voltage, as shown in Fig. 2.3. The actual conversion starts after driving all bottom plates to ground, with the top plates allowed to float.

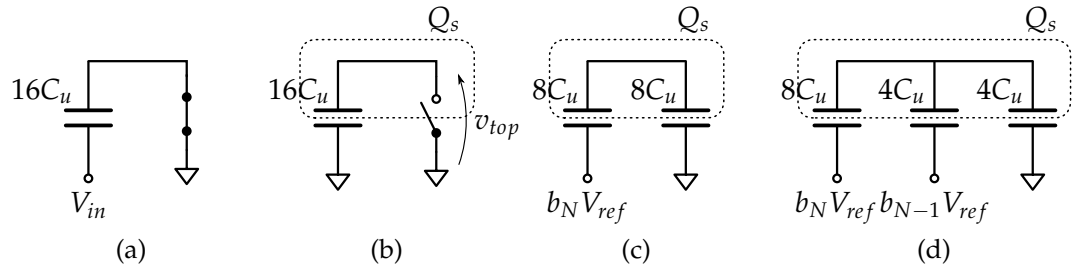

Figure 2.4.: Capacitive array during: a-b) sampling, c-d) conversion. The charge in the isolated node remains constant across the operations.

Using the notation $v[n] \stackrel{def}{=} v(nt_{clk})$, where n represents the conversion cycle and t_{clk} the

time required by each step, the initial array voltage can be expressed as

$$v_{top}[0] = -V_{in}$$

and the most significant bit b_N, considering a bipolar representation where $b_i \in \{-1, +1\}$, is obtained as

$$b_N = \text{sign}\left(v_{top}[0] - V_{cm}\right).$$

In the circuit in Fig. 2.3 V_{cm} corresponds to the ground potential. If v_{top} is lower than the common mode level, $b_N = -1$ and the array voltage has to be increased to get closer to the common mode. The value of b_N determines the connection of the MSB capacitor to either V_{ref}^{+} or V_{ref}^{-} resulting in a new array voltage. This value can be derived by considering the fixed charge stored on the top plates and the new connection of the largest capacitor.

The array can be grouped into two elements, $C_{MSB} = 8C_u$ and $C_{rem} = 8C_u$, such that $C_{MSB} + C_{rem} = C_{tot} = 16C_u$ (Fig. 2.4c). Forcing the conservation of charge, we obtain

$$-V_{in}C_{tot} = \left(v_{top}[1] - b_N V_{ref}\right) C_{MSB} + C_{rem} v_{top}[1]$$

and consequently

$$v_{top}[1] = -V_{in} + b_N \frac{C_N}{C_{tot}} V_{ref} = -V_{in} + \frac{b_N}{2} V_{ref}.$$

Since the MSB capacitor represents half the total capacitance, the array voltage changes by half the reference voltage.

For the second bit:

$$-V_{in}C_{tot} = \left(v_{top}[2] - b_N V_{ref}\right) C_{MSB} + \left(v_{top}[2] - b_{N-1} V_{ref}\right) C_{MSB-1} + C_{rem} v_{top}[2].$$

Coherently with the fact that the (MSB-1) capacitor has one fourth of the array capacitance, the voltage becomes

$$v_{top}[2] = -V_{in} + \left(\frac{b_N}{2} + \frac{b_{N-1}}{4}\right) V_{ref}.$$

As conversion proceeds, new bits are generated, each of them determining the position of one selector. Since the capacitors have values that decrease as powers of two, every new bit leads to an increasingly smaller variation of the array voltage. At the end of the conversion

$$\begin{aligned} v_{top}[N] &= -V_{in} + \frac{b_N C_N + b_{N-1} C_{N-1} + \cdots + b_0 C_0}{C_{tot}} V_{ref} \\ &= -V_{in} + \left(\frac{b_N}{2} + \frac{b_{N-1}}{4} + \cdots + \frac{b_0}{2^{N+1}}\right) V_{ref}. \end{aligned} \tag{2.1}$$

The fact that the number of terms in (2.1) is $N+1$ whereas the capacitors are N stems from the fact that the first bit is generated by grounding all the array elements, and the remaining N by acting on the capacitors (the closure capacitance of value C_u is excluded from the count since it is introduced only to have power-of-two coefficients and is unused during conversion). The quantization error is therefore bounded by

$$|\varepsilon_q| < \frac{1}{2^{N+1}} V_{ref}.$$

As N is increased, the error is reduced and the approximation improves. However, every new bit introduces a capacitance equal to the one of the entire array, with an exponential increase of both the area and the power consumption. The total capacitance is in fact

$$C_{tot} = 2^N C_u.$$

As a consequence, also the impedance of both the source and the switches has to be extremely low to avoid slow transients.

2.2.2. Split Array and C-2C

The greatest concern as resolution is increased is matching of the capacitive elements, since the operations of the DAC depend inescapably on exact ratios of capacitance. As larger capacitors are added to introduce new bits, guaranteeing the accuracy of the ratios becomes problematic. Therefore alternative topologies have to be considered to increase the resolution. Two structures typically employed are the split array and the C-2C sub-array (Fig. 2.5).

In the scaled capacitive array considered in the previous section, all the capacitors shared the top node. If a series element (bridge) is introduced, the array is split. Between the two top nodes now present, only one needs to be grounded during sampling and subsequently observed by the comparator. In Fig. 2.5a it is the left node, since the closure element C_u is on the right-hand side.

Looking from the comparator input, the capacitance on the other side of the bridge is attenuated, acting as an equivalent smaller elements. However, the capacitances in the secondary array have values already used in the main one. The range of capacitances is thus limited and matching can be guaranteed more easily. Eventually the entire primary array, having N_p bits could be replicated, doubling the resolution while doubling the area and total capacitance $2^{N_p+1}C_u + C_b$. The same resolution increase, using exclusively a scaled array, would require a capacitance $2^{2N_p}C_u$, the square of the original one.

The value of bridge capacitance making the secondary array look as an extension of the scaled array is derived by considering that its total capacitance $C_{sec} = 2^{N_s}C_u$ in series with C_b has to equal C_u (to have a total capacitance, as seen from the input, expressed as a power of two). Therefore

$$C_b = \frac{2^{N_s}}{2^{N_s} - 1} C_u.$$

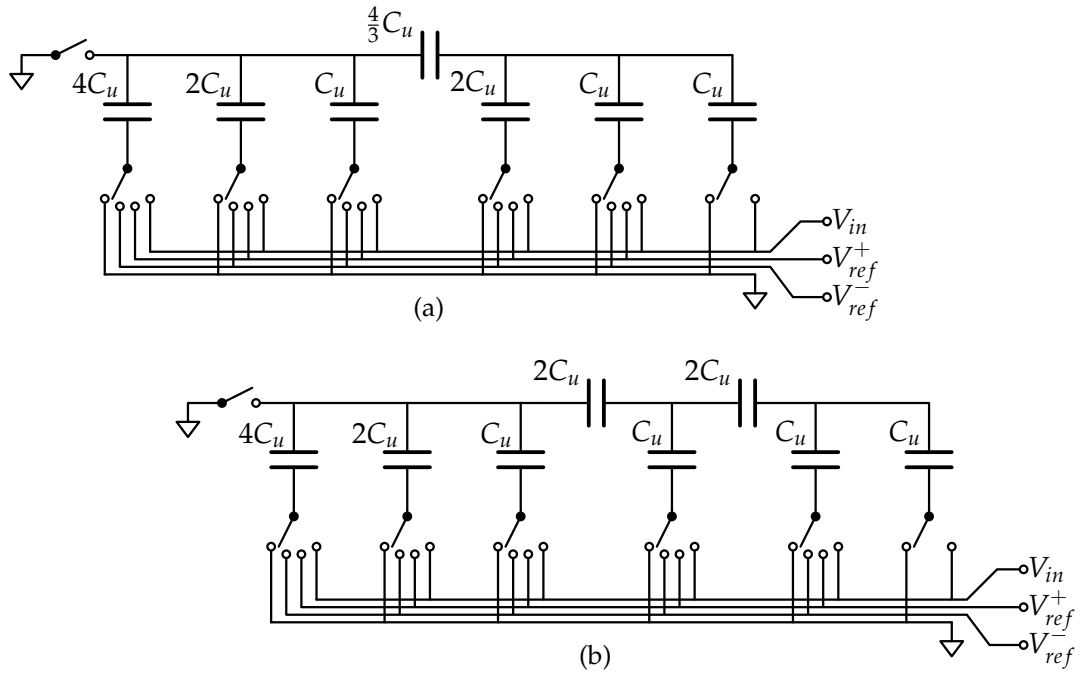

Figure 2.5.: a) Split array with unequal sections b) Mixed-type array: scaled plus C-2C sections. The comparator in both cases would be connected to the same node as leftmost switch.

The values of bridge capacitance are not integer multiples of the C_u, requiring a careful layout. The procedure could be repeated even more than once, adding several bridges. However, having non-integer multiples of C_u and introducing at the same time many isolated nodes reduces the achievable accuracy, making the structure extremely sensitive to injected noise and parasitic loading, with detrimental effects on the conversion quality.

Notwithstanding this last consideration, the structure that minimizes the total capacitance and that has been actually considered in the proposed solution is the mixed-type array, employing a C-2C topology. It involves the cascade of several capacitive dividers, with values $2C_u$ and C_u and loading the original scaled array. One cell per each new bit is introduced, with a final closure capacitance equal to C_u. In Fig. 2.5b, two C-2C cells are added to the original scaled array. The total capacitance expressed as a function of the number of bits in the sub-array is:

$$C_{C2C} = (3N + 1)C_u$$

This is lower than in a scaled implementation if $N \geq 4$. Moreover, requiring only two values of capacitance, high matching can be easily achieved.

The downside of this solution is the presence of the many isolated nodes. What will be shown in Section 3.2 is that the amount of parasitics loading the internal nodes

determines the maximum number of bits achievable with such a structure. This is the main limitation preventing the realization of the entire array as a cascade of C-2C cells.

Evaluation of the array voltage in both the split and C-2C solutions, as already done in (2.1) for the scaled topology, is slightly more complicated. However, using superposition and considering the expression of a capacitive voltage divider under the assumption of no net charge in the hidden isolated nodes, the result can be obtained. Indeed the expressions in the case of the C-2C-based mixed-type array will be derived in Section 3.1.1.

2.3. Capacitive Array for CS-based Acquisitions

The focus of this section is on the modifications to be applied to the mixed-type capacitive array (Fig. 2.5b) in order to implement the linear projection $y = Sx$ involved in a CS-based acquisition. Component-wise, the product can be formulated as:

$$y_j = \sum_{k=1}^{n} S_{j,k} x_k \qquad j = 1, \ldots, m, \tag{2.2}$$

where $S_{j,k}$ is the element at the intersection of the j-th row and k-th column of the sensing matrix, x_k is a signal sample and y_j is a measurement. Equation (2.2) requires m replicas of a multiply-and-accumulate stage. The product between matrix elements and signal samples can be trivially implemented only if the coefficients belong to $\{-1, 0, 1\}$. In a differential implementation, in fact, multiplication by -1 only requires the inversion of the signal polarity, easily manageable with four switches in total. The modulation coefficients can be provided from an external source, can be generated in advance and stored in an on-chip memory, or created using a pseudo-random generator followed by a proper filter that shapes the correlation of the terms. Refer to [5] for how to obtain sequences with specific second order statistics.

Concerning the summation of the modulated samples, one approach could be the use of a switched-capacitor, discrete-time integrator, as already done in previous implementations like [11]. The major drawback is the power consumption of the operational amplifier required by each integrator. Moreover, the voltage can saturate for some sequences of input values and modulating coefficients, leading to a loss of information on the measurements unless some ad-hoc technique is employed. Both these issues are avoided by the use of a completely passive solution.

Consider what happens when n identical capacitors of value C_s, precharged at different voltages v_k, are later connected in parallel. The voltage across all of them can be expressed as:

$$v = \frac{Q_{tot}}{C_{tot}} = \frac{1}{nC_s} \sum_{k=1}^{n} C_s v_k = \frac{1}{n} \sum_{k=1}^{n} v_k \tag{2.3}$$

We get the average of the initial voltages. Apart from the scaling factor $1/n$, and if we consider each v_k as a modulated input $S_{j,k}x_k$ sample, this is exactly what (2.2) requires. This behaviour can be obtained from the capacitive array if the capacitors do not sample the input at the same time. The array has to be decomposed into elements of uniform capacitance, each acquiring a modulated sample at different instants.

Starting from the original structure of the array, depicted in Fig. 2.6 (top), this means that the total capacitance C_{tot} has to be decomposed into elements of value $C_h = C_{tot}/n$. If n is a power of two, C_h is a multiple of the unitary capacitance. In general, having $C_u < C_h < C_{tot}$, the MSB capacitors have to be decomposed into smaller elements, while the LSB ones are driven together, sampling at the same instant and acting effectively as a larger element. Fig. 2.6 shows exactly how the $4C_u$ element is split in half, while the entire C-2C array, which from the perspective of the input behaves as a capacitance of value C_u, is collected with the smallest scaled capacitor to form a sampling element of value $2C_u$.

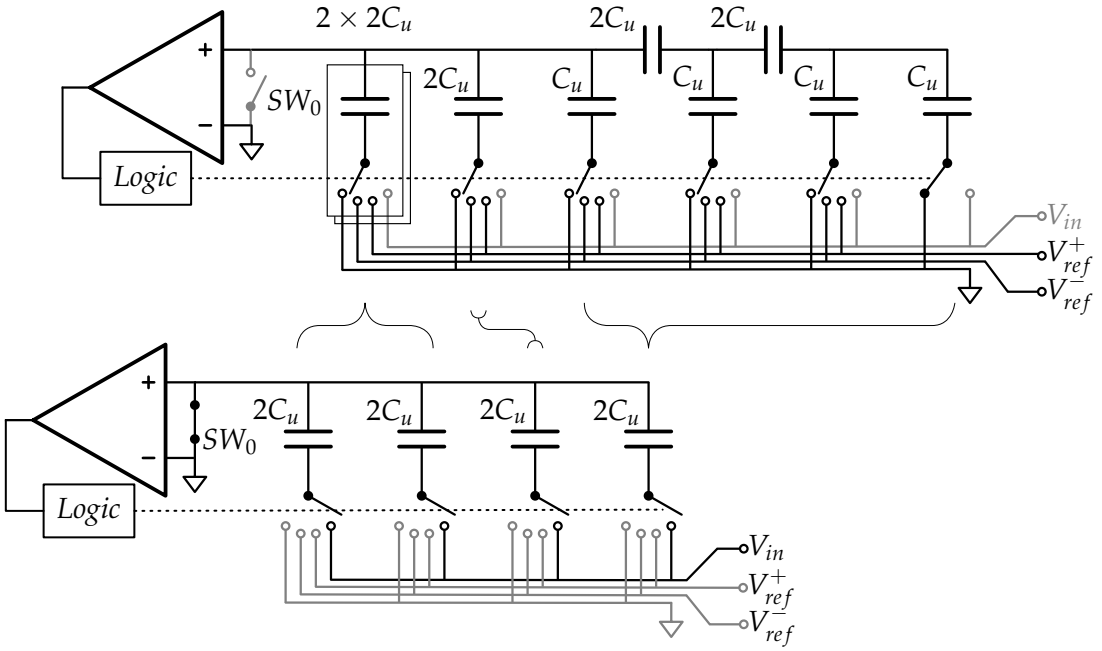

Figure 2.6.: A/I converter employing a mixed-type capacitive array during conversion (top) and sampling (bottom).

The acquisition now lasts for an entire signal window, i.e. n samples are collected at Nyquist rate. At each sampling instant, one capacitive hold element stores the modulated signal. At the end of the window, the top plate is left floating (SW_0 opened) and the bottom plates are grounded (through the selectors). The array voltage observed by the comparator becomes the average of the sampled values, as in (2.3). Conversion then proceeds as usual, with the sub-elements belonging to a large capacitor driven

together so that they behave as the original component. Apart from a modification in the timing of the control signals, the only addition is new selection switches to each capacitor to be decomposed ($4C_u$ in the example). This is indeed why this solution looks promising. There are no extra active elements affecting the overall power budget.

In the example shown in Fig. 2.6, the C-2C array has not been decomposed. It is the preferred solution since this kind of structure has very sensitive internal nodes, which are better left untouched. Only $C_h \geq C_u$ is then possible. Therefore the entire array is always considered, as seen from the input, as a unique capacitance of value C_u. If $C_h = C_u$ the array is driven as an independent element. If instead $C_h > C_u$ then it is combined with some of the smallest scaled capacitances to form one entire hold capacitor.

2.3.1. Acquisition with leaky elements

The converter in Fig. 2.6 implements one row of the sensing matrix, therefore, if the matrix $S \in \mathbb{R}^{m \times n}$ is full, m converters are required, each with the ability to decompose its internal array into n elements. With a CR in the range $[2 : 16]$ and $n = 256$, the number of channels m will be comprised in $[16 : 128]$.

A large n leads to the decomposition of the capacitors into extremely fine elements, leading to particularly small values of the hold capacitance. At the same time, samples have to be preserved until the end of an acquisition window, which for ECG-type signals is around one second long. Since capacitors discharge over time because of leakage

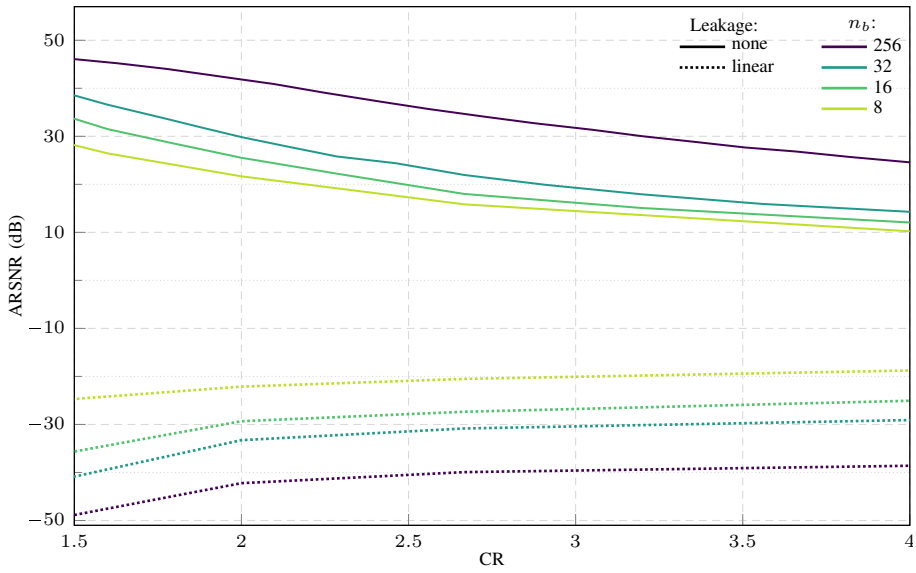

Figure 2.7.: ARSNR after leakage

currents induced by the sorrounding elements, we expect that using a full matrix is

unfeasible. This is indeed what the curves in Fig. 2.7 show: all the information is lost before conversion and the result from reconstructing the original waveform is only noise (ARSNR < 0).

The curves are obtained assuming the discharge of the hold capacitors by the reverse saturation currents of the switches' junctions. Using data from a commercial 180 nm technology, the leakage current due to a set of minimum size switches has been estimated to be 300 pA. As an example, starting from an initial voltage of 1.8 V a hold capacitor of 1 pF is completely discharged after 6 ms, much less than the length of an ECG window, which spans around 1 s.

Using a block diagonal sensing matrix, the A/I converter works on $n_b < n$ samples. Moreover, keeping a constant C_{tot} both to have the same power consumption and area occupation of a single converter, each hold capacitor is now larger. The combination of a shorter window and a larger capacitance leads, as shown by the lighter curves in Fig. 2.7 to a higher reconstruction quality. However we still have ARSNR < 0, i.e. mainly noise.

This result is the reason behind the introduction of a compensation loop around every hold capacitor, so that the acquired samples can be preserved until the end of the acquisition window, which, in any case, should not be excessively long.

Part II.

Modeling and Design

3. Limitations of the C-2C Array

As described in Chapter 2.3, the capacitive array is at the core of the proposed A/I converter. It is used to collect and store several modulated samples of the input signal, combine them and convert the result. All these operations rely on the redistribution of the charge stored in the isolated nodes.

To increase the resolution of the conversion, we have shown that the most efficient solution is to cascade a C-2C sub-array to the smallest scaled capacitor. This structure, however, introduces secondary isolated nodes which are particularly sensitive to injected noise and parasitic loading. Figure 3.1 shows a 3-bit C-2C structure loading a scaled array, which has been compacted into one single capacitance C_{sc}. In this section, the smallest element of the scaled array is considered as part of the C-2C structure. Each isolated node has been named as N_i and all the parasitics have been highlighted.

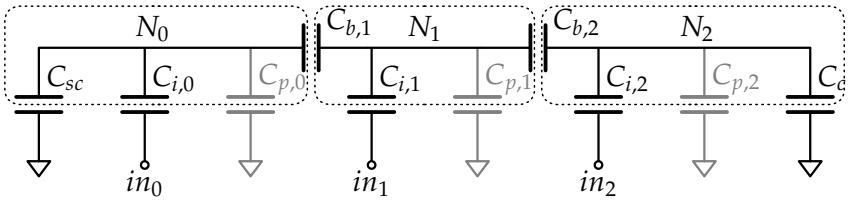

Figure 3.1.: Schematic of a 3-bit C-2C sub-array. The elements are (nominal value in parenthesis): input capacitors C_i (C_u); bridge capacitors C_b ($2C_u$); parasitic capacitors C_p (0); equivalent capacitance of the scaled array C_{sc} ($2C_u + 4C_u + \cdots$); closure capacitance C_c (C_u)

In order to evaluate the maximum depth of the C-2C structure, the effects of parasitics and mismatch have been analyzed, obtaining a design constraint on the maximum extension of the sub-array i.e. the number of bits.

3.1. Complete model

Since the SAR algorithm evolves according to the voltage of the main isolated node N_0, the effect of any non-ideality has to be observed from such node.

Consider each capacitor described by its nominal value plus an error term due to imperfections. The error term is called ε for the input capacitors and ε_b for the bridge capacitors. Furthermore the metal plates used to realize the devices couple to the external environment, resulting in some parasitic loading of the inner nodes, here expressed

as a fraction α of the unit capacitance. The parameters involved in the analysis are summarized below:

$$\begin{aligned}
n &: \text{inputs of the C-2C array} & C_{sc} &= 2C + 4C + \cdots \text{ (scaled capacitance)} \\
i &: \text{section index, } [0 : n-1] & C_{i,i} &= C_u(1+\varepsilon_i) \\
N_i &: \text{isolated nodes} & C_c &= C_u(1+\varepsilon_n) \\
in_i &: \text{input nodes} & C_{b,i} &= 2C_u(1+\varepsilon_{b,i}) \\
n_{sc} &: \text{scaled } (C_{sc}) \text{ inputs} & C_{p,i} &= \alpha C_u
\end{aligned} \tag{3.1}$$

A note on the n_{sc} parameter. It represents the number of scaled capacitors, if present, starting from the term $2C_u$ (the C_u capacitor now is considered part of the C-2C sub-array). If the C-2C structure works in isolation, then C_{sc} and n_{sc} are both zero. The effects of a voltage applied to one input of the C-2C sub-array depends on the amount of scaled capacitance loading node N_0, thus the need to consider it.

As a voltage is applied to the input nodes, the variation observed at N_0 depends on the topological distance of the input from N_0, coherently with the fact that each input represents a different bit position. In the presence of non-idealities, the attenuation will be affected by the errors of all the components in the array, each weighted by some coefficient depending on the distance from the nodes. Only by deriving the exact equations describing the structure each single effect can be quantified.

3.1.1. Nominal bit weights

The initial analysis will lead us to the nominal value of the attenuation $R_{i,0}$ from a generic input in_i to the main isolated node N_0, so as to determine what the comparator will observe when a voltage variation is applied at the input terminals. The attenuation can be decomposed in two terms. The first considers the capacitive partition from in_i to N_i, the second from N_i to N_0. This, in turn, is due to the cascade of attenuators from N_i to N_{i-1} to N_{i-2} until N_0. Therefore:

$$R_{i,0} \stackrel{\text{def}}{=} R_{in_i,N_i} R_{N_i,N_{i-1}} R_{N_{i-1},N_{i-2}} \cdots R_{N_1,N_0}. \tag{3.2}$$

The attenuations between adjacent isolated nodes and from the input to its closest inner node are shown in Table 3.1 and Table 3.2.
The symbol $\oplus$ describes the harmonic sum of the two operands, defined as:

$$x \oplus y \stackrel{\text{def}}{=} \frac{1}{\frac{1}{x}+\frac{1}{y}} = \frac{xy}{x+y}$$

and refers to the series connection of capacitors.

The overall attenuation $R_{i,0}$ defined in (3.2) is given by the product of all the rows in Table 3.1 starting from the first, up to the one of $R_{N_i,N_{i-1}}$ times the row of R_{in_i,N_i} in

Table 3.1.: Attenuation felt by a voltage as it propagates from one isolated node to the next, towards N_0. Results shown for the first 3 nodes.

Attenuation	Definition	Result
R_{N_1,N_0}	$\frac{2C}{2C+C+C_{sc}}$	$\frac{2}{1+2^{n_{sc}+1}}$
R_{N_2,N_1}	$\frac{2C}{2C+C+2C\oplus(C+C_{sc})}$	$2\frac{1+2^{n_{sc}+1}}{1+5\cdot 2^{n_{sc}+1}}$
R_{N_3,N_2}	$\frac{2C}{2C+C+2C\oplus(C+2C\oplus(C+C_{sc})}$	$2\frac{1+5\cdot 2^{n_{sc}+1}}{1+21\cdot 2^{n_{sc}+1}}$

Table 3.2.: Attenuation felt by the input voltage, observed from the closest isolated node. Results shown for the first 4 inputs.

Attenuation	Definition	Result
R_{in_0,N_0}	$\frac{C}{2C+C_{sc}}$	$\frac{1}{2^{n_{sc}+1}}$
R_{in_1,N_1}	$\frac{C}{C+C+2C\oplus(C+C_{sc})}$	$\frac{1+2^{n_{sc}+1}}{2^{n_{sc}+3}}$
R_{in_2,N_2}	$\frac{C}{C+C+2C\oplus\left(C+2C\oplus(C+C_{sc})\right)}$	$\frac{1+5\cdot 2^{n_{sc}+1}}{2^{n_{sc}+5}}$
R_{in_3,N_2}	$\frac{C}{C+C+2C\oplus\left(C+2C\oplus\left(C+2C\oplus(C+C_{sc})\right)\right)}$	$\frac{1+21\cdot 2^{n_{sc}+1}}{2^{n_{sc}+7}}$

Table 3.2. In the first table, the numerator of each row cancels the denominator of the previous one, leaving as a result 2^i divided by the last denominator. In turn, this term is exactly equal to the numerator of the corresponding row in Table 3.2, leading to

$$R_{i,0} = \frac{2^i}{2^{n_{sc}+2i}} = \frac{1}{2^{n_{sc}+i}} \tag{3.3}$$

This proves the fact that the weights of the C-2C inputs decrease as powers of two (dependence on $1/2^i$), with the absolute attenuation depending on how big is the scaled array. Including all the non-idealities would allow us to obtain the exact equations describing the behaviour of the structure. Since this process is tedious and error prone, an algorithmic solution has been preferred.

3.1.2. Algorithmic derivation of the complete model

To derive the array equations for an arbitrary number of inputs, the model depicted in Figure 3.2 has been analyzed. It decomposes the equivalent capacitance seen from one node into the left-hand component C_L, the right-hand one C_R, and the parasitic C_p, input C_i and bridge C_b capacitances. These last three terms are represented with their respective errors ε, ε_b and α (C_p by itself is an unwanted component, entirely).

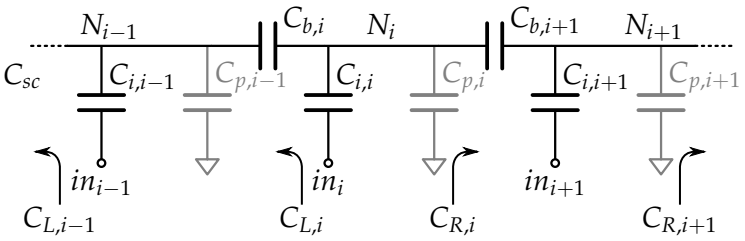

Figure 3.2.: Capacitances in a generic section of a C-2C array

The equations describing one section are:

$$\begin{aligned} C_{L_i} &= (C_{L_{i-1}} + C_{in_{i-1}} + C_{p_{i-1}}) \oplus C_{bi} \\ C_{R_i} &= (C_{R_{i+1}} + C_{in_{i+1}} + C_{p_{i+1}}) \oplus C_{b_{i+1}} \\ R_{N_i,N_0} &= R_{N_i,N_{i-1}} \frac{C_{b_i}}{C_{b_i} + C_{L_{i-1}} + C_{in_{i-1}} + C_{p_{i-1}}} \\ R_{in_i,N_0} &= R_{N_i,N_0} \frac{C_{in_i}}{C_{in_i} + C_{L_i} + C_{R_i} + C_{p_i}} \end{aligned}$$

The model is recursive, since the relationships at one node are defined on what happens at the adjacent node, with the elementary conditions:

$$\begin{aligned} C_{L_0} &= C_s \\ C_{R_n} &= C_c \end{aligned}$$

The pseudocode describing how to evaluate every term is shown below. Any symbolic manipulation library can be used to implement it (e.g. `sympy` for Python).

```
Initialize C_in_i, C_b_i and C_p_i according to (3.1)
C_L_0 = C_s
C_R_n-1 = C_c
R_N_0,N_0 = 1
for i = n-2:0
   compute C_R_i
for i = 1:n-1
   compute C_L_i
   compute R_N_i,N_0
   compute R_in_i,N_0
   cancel second order terms
```

The resulting attenuations will be expressed as a fraction whose numerator and denominator are first order approximations of the real ones. Assuming small errors (ε, ε_b and α), as desirable, this is a reasonable approximation. Each component will thus be expressed

as the nominal value plus three error terms, the first due to mismatches (ε and ε_b), the last to parasitics (α):

$$R_{i,0} = \frac{v_{N_0}}{v_{in_i}} \approx \frac{1}{2^{N_s+i}} \frac{1 + \sum_{j=0}^{n} \left(k_{i,j}^{\varepsilon} \varepsilon_j + k_{i,j}^{\varepsilon_b} \varepsilon_{bj} + k_{i,j}^{\alpha} \alpha_j \right)}{1 + \sum_{j=0}^{n} \left(h_{i,j}^{\varepsilon} \varepsilon_j + h_{i,j}^{\varepsilon_b} \varepsilon_{bj} + h_{i,j}^{\alpha} \alpha_j \right)} \tag{3.4}$$

The coefficients $k_{i,j}$ and $h_{i,j}$ represent the weight by which every non-ideality in position j in the array affects the result, when the input is applied at i. Since the error components have different nature we will deal with them separately.

3.2. Effects of parasitics

The first order approximation of the attenuations from inputs in_i to the main isolated node N_0, considering only the effects of parasitics and for a few sizes of the array are shown in Table 3.3. X represents the total scaled capacitance, normalized with respect to the unitary capacitance, i.e. $X = C_{sc}/C_u = 2^{n_{sc}+1} - 2$.

Table 3.3.: First order approximations of the attenuations from input i to node N_0 in the case of (a) $n = 2$, (b) $n = 3$ and (c) $n = 4$

in	R_{in_i,N_0}
0	$\frac{4+\alpha_1}{8+4X+4\alpha_0+3\alpha_1+X\alpha_1}$
1	$\frac{2}{8+4X+4\alpha_0+3\alpha_1+X\alpha_1}$

(a)

in	R_{in_i,N_0}
0	$\frac{16+4\alpha_1+5\alpha_2}{32+16X+16\alpha_0+12\alpha_1+11\alpha_2+4X\alpha_1+5X\alpha_2}$
1	$\frac{8+2\alpha_2}{32+16X+16\alpha_0+12\alpha_1+11\alpha_2+4X\alpha_1+5X\alpha_2}$
2	$\frac{4}{32+16X+16\alpha_0+12\alpha_1+11\alpha_2+4X\alpha_1+5X\alpha_2}$

(b)

in	R_{in_i,N_0}
0	$\frac{64+16\alpha_1+20\alpha_2+21\alpha_3}{128+64X+64\alpha_0+48\alpha_1+44\alpha_2+43\alpha_3+16X\alpha_1+20X\alpha_2+21X\alpha_3}$
1	$\frac{32+8\alpha_2+10\alpha_3}{128+64X+64\alpha_0+48\alpha_1+44\alpha_2+43\alpha_3+16X\alpha_1+20X\alpha_2+21X\alpha_3}$
2	$\frac{16+4\alpha_3}{128+64X+64\alpha_0+48\alpha_1+44\alpha_2+43\alpha_3+16X\alpha_1+20X\alpha_2+21X\alpha_3}$
3	$\frac{8}{128+64X+64\alpha_0+48\alpha_1+44\alpha_2+43\alpha_3+16X\alpha_1+20X\alpha_2+21X\alpha_3}$

(c)

Assuming $\alpha_j = \alpha$ for each j, the numerical coefficients can be collected and their patterns exploited. The overall contribution can be expressed by a simple formula

dependent only on n and i, resulting in a description of the effects of parasitics for a generic size of the array and from any input, just as desired. Expressing each fraction in Table 3.3 as

$$R_{in_i,N_0} = \frac{A_{nom,i} + \alpha A_{\alpha,i}}{B_{nom,i} + \alpha B_{\alpha,i} + X(C_{nom,i} + \alpha C_{\alpha,i})},$$

where A, B and C collect the numerical coefficients in the fraction, the relationships among the numbers in the original expression have to be unveiled. The nominal terms are equal to

$$A_{nom,i} = \frac{4^{n-1}}{2^i}$$

$$B_{nom,i} = 2 \cdot 4^{n-1}$$

$$C_{nom,i} = 4^{n-1}$$

resulting in the nominal value of the attenuation already found in (3.3):

$$R^{nom}_{in_i,N_0} = \frac{A_{nom,i}}{B_{nom,i} + X \cdot C_{nom,i}} = \frac{1}{2+X}\frac{1}{2^i} = \frac{1}{2^{n_{sc}+1}}\frac{1}{2^i}.$$

Concerning the coefficients of the error, we get:

$$\begin{aligned}
A_{\alpha,i} &= \frac{4^{n-2}}{2^i}\sum_{j=0}^{n-i-2}\frac{1}{4^j}(n-i-j-1) = \frac{2^i}{9}\left[1 + 4^{n-i}\left(\frac{3}{4}(n-i)-1\right)\right] \\
B_{\alpha,i} &= -\sum_{j=1}^{n-1} j4^{j-1} + n4^{n-1} \qquad = \frac{1}{9}\left[\left(\frac{3}{2}n+1\right)4^n - 1\right] \\
C_{\alpha,i} &= \sum_{j=1}^{n-1} j4^{j-1} \qquad = \frac{1}{9}\left[\left(\frac{3}{2}n-1\right)4^n + 1\right]
\end{aligned} \tag{3.5}$$

Using Taylor's expansion of R_{in_i,N_0}:

$$\begin{aligned}
R_{in_i,N_0} &\simeq R^{nom}_{in_i,N_0}\left(1 + \frac{\alpha A_{\alpha,i}}{A_{nom,i}}\right)\left(1 - \frac{\alpha B_{\alpha,i}}{B_{nom,i} + XC_{nom,i}} - \frac{\alpha XC_{\alpha,i}}{B_{nom,i} + XC_{nom,i}}\right) \\
&= R^{nom}_{in_i,N_0}\left(1 + \frac{\alpha A_{\alpha,i}}{A_{nom,i}} - \frac{\alpha B_{\alpha,i}}{B_{nom,i} + XC_{nom,i}} - \frac{\alpha XC_{\alpha,i}}{B_{nom,i} + XC_{nom,i}}\right)
\end{aligned}$$

the coefficients at the numerator and denominator can be combined through their relative value. Thus, substituting (3.5) and performing some simplifications considering, reasonably, $n \geq 2$ and $n - i \geq 2$, we find:

$$R_{I_i,N_0} = \frac{1}{2+X}\frac{1}{2^i} - \frac{1}{2+X}\frac{1}{2^i}\frac{\alpha}{3}\left(i + \frac{n}{2X}\right) \tag{3.6}$$

$$= \frac{1}{2+X}\frac{1}{2^i}\left(1 - \frac{\alpha}{3}\left(i + \frac{n}{2X}\right)\right) \tag{3.7}$$

The last inequality ($n - i \geq 2$), required to get a closed form solution, implies that the approximation does not hold for the last input. This is sufficient for our purposes, especially since the effects we want to observe and will actually determine the size of the array refer to the most significant inputs.

It is worth to observe that the relative error (3.7) in a given structure (fixed n and X) increases linearly with the input i. At the same time, since the nominal value of the weight decreases exponentially, the absolute error decreases with the position of the input. Therefore, a constraint can be defined on the absolute error of input $i = 1$, comparing it against an LSB, in order to force the error induced by parasitics to be lower than some threshold.

Consider the total number of bits of the converter:

$$n_{bit} = n_{sc} + n + 1 = \log_2(2 + X) + n,$$

one LSB is evaluated as

$$LSB = \frac{1}{2^{n_{bit}}} = \frac{1}{(2 + X)2^n}. \tag{3.8}$$

The quantity we are interested in is the maximum absolute error (rightmost addend in (3.7)) normalized in terms of one LSB. This should be lower than some threshold, expressed here as $1/2^l$ just so that the final expression is clearer (e.g. for $l = 2$ the error we tolerate is LSB$/2^l$ =LSB$/4$). Since the error decreases for $i > 0$, the maximum is found for $i = 1$, therefore:

$$\max\left(\frac{\varepsilon_{abs}}{\text{LSB}}\right) = \frac{\alpha}{6}2^n\left(1 + \frac{n}{2X}\right) < \frac{1}{2^l}$$

Moving α on the other side of the inequality, we obtain a quantity ε_{norm} that depends only on the topology of the array (n and X), which has to be lower than a threshold defined by the amount of parasitics α and a design parameter l:

$$\varepsilon_{norm} = \frac{2^n}{6}\left(1 + \frac{n}{2X}\right) < \frac{1}{\alpha 2^l}$$

The behaviour of ε_{norm} is plotted in Fig. 3.3. In the case of large X ($X > 5n$) the second term in the parenthesis can be neglected, obtaining a closed form expression for the bound on n:

$$n < 2.585 - \log_2 \alpha - l$$

For smaller X, the bound is lower, as shown in the figure. The intersection of said curves with $1/(\alpha 2^l)$ determines the feasible widths of the array. As an example, with $\alpha = 0.1$ and $l = 2$ (maximum error of LSB$/4$), n should be between lower than 4, depending on the amount of scaled capacitance.

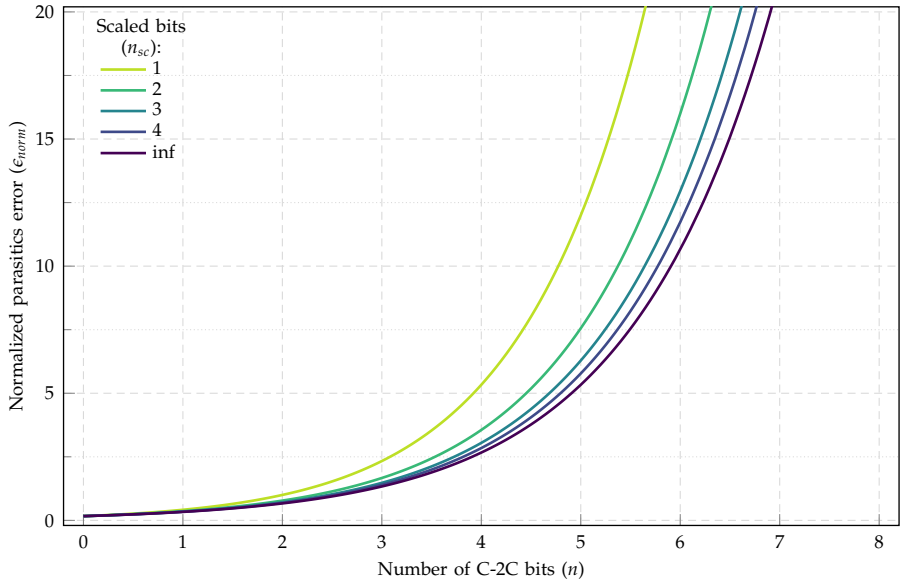

Figure 3.3.: Maximum error due to the C-2C parasitics, normalized with respect to α and one LSB. Curves are parameterized in terms of the size of the scaled array.

3.2.1. Nonlinearities

So far we have derived the error on each single bit realized by a C-2C structure affected by parasitic loading at its inner isolated nodes. This information can be used to derive the integral and differential nonlinearities, which describe how the real A/D conversion performs with respect to an idealized one. The parameters are defined as follows [10]:

$$\mathrm{INL}(x) \stackrel{\mathrm{def}}{=} \frac{A(x) - x\,\mathrm{LSB}}{\mathrm{LSB}} \tag{3.9}$$

$$\mathrm{DNL}(x) \stackrel{\mathrm{def}}{=} \frac{A(x+1) - A(x)}{\mathrm{LSB}} - 1 \tag{3.10}$$

where x is the digital code being represented and $A(x)$ the corresponding analog value. Both parameters are normalized with respect to one ideal step (LSB). Let us first express the INL as a function of the error associated to each bit. The term $A(x)$ in (3.10) is substituted with (3.7) and one LSB with (3.8), obtaining:

$$\begin{aligned}\mathrm{INL}(x) &= \frac{\sum_{i=0}^{n-1} b_i R_{in_i,N_0} - x\,\mathrm{LSB}}{\mathrm{LSB}} = \frac{\sum_{i=0}^{n-1} b_i \varepsilon_{par,i}}{\mathrm{LSB}} \\ &= \sum_{i=0}^{n-1} \frac{\alpha 2^n}{3}\frac{b_i}{2^i}\left(i + \frac{n}{2X}\right) \simeq \sum_{i=0}^{n-1} \frac{\alpha 2^n}{3}\frac{i b_i}{2^i}\end{aligned} \tag{3.11}$$

where the the value (0 or 1) of the single bits b_i, depends on x. The values of $\mathrm{INL}(x)$ for different number of C-2C bits n and normalized with respect to α are plotted in Fig. 3.4.

Since (3.11) is a summation of nonnegative terms, the maximum value is reached for sure on the last element, i.e. when $b_i = 1$ in every position. In that case, $x = 2^n - 1$ and the sum becomes:

$$\max_{0 \le x < 2^n} \left[\mathrm{INL}(x)\right] = \mathrm{INL}(2^n - 1) = \frac{2\alpha}{3}\left(2^n - (n+1)\right).$$

This formula gives us the maximum INL as a function of the number of bits and the amount of parasitics. It determines the effective number of bits of the converter, when limited only by the parasitics.

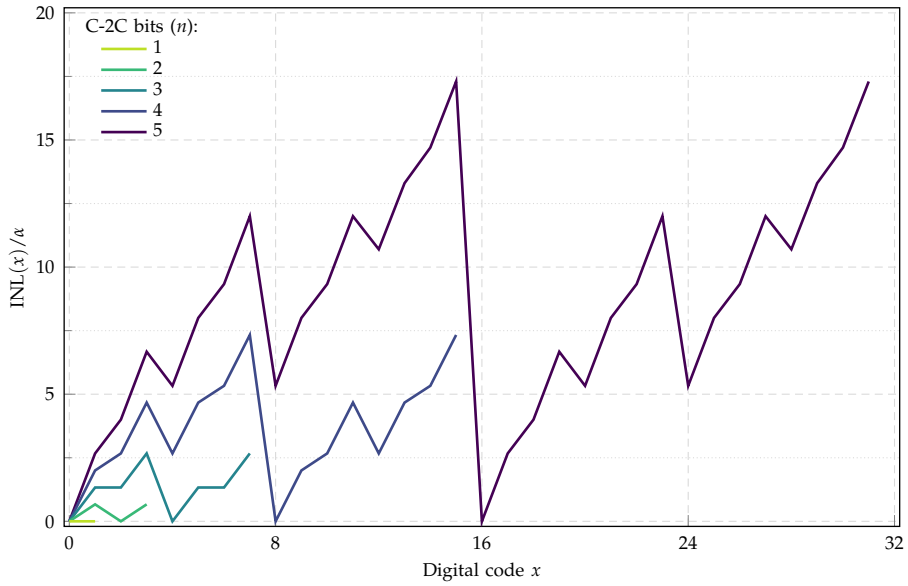

Figure 3.4.: INL normalized with respect to α for several widths of the C-2C array.

Being always positive, the INL could be reduced by a proper characterization of the converter, defining, as an example, the best-fit line minimizing the overall quadratic error.

The DNL compares the local variation between consecutive steps to one LSB. It can also be expressed as $\mathrm{INL}(x+1) - \mathrm{INL}(x)$, therefore looking at Fig. 3.4 we readily observe that the DNL is maximum (in absolute value) in correspondence of the halfway transition. This stems from the fact that the sequence goes from '01…1' to '10…0', the former combines the error of all the bits other than the first, the latter only shows the contribution of the first. Fig. 3.5 depicts the DNL corresponding to the curves in Fig. 3.4.

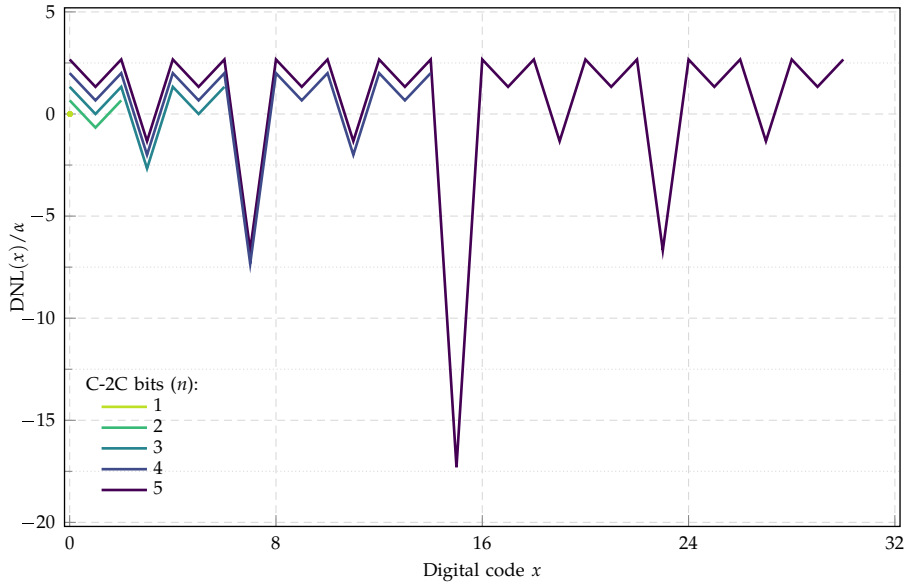

Figure 3.5.: DNL normalized with respect to α for several widths of the C-2C array.

The maximum DNL can be expressed as:

$$\max_{0 \leq x < 2^n} \left|\mathrm{DNL}(x)\right| = \mathrm{DNL}(2^{n-1} - 1) = -\frac{2\alpha}{3}\left(2^n - (n+1)\right)$$

If α is sufficiently large, the negative peaks might cross the $\mathrm{DNL} = -1$ threshold that results in a non-monotonic D/A characteristic.

3.3. Mismatch considerations

Mismatch errors on each capacitance can be considered as identically-distributed random variables, with zero mean and equal variance. An intuitive understanding of what happens in a C-2C array when the mismatch among the elements has unitary correlated comes from the observation that if all capacitances grow by some factor, say $1 + k$, than the array should still behave as a C-2C, since $C(1 + k)$ and $2C(1 + k)$ still have a ratio of 2. The entire structure is then insensitive to mismatch, which is unreasonable.

Further work is required to evaluate reasonable bounds on the unitary capacitance given the statistics of the technological process.

4. Switch configuration

In any sample and hold circuit, it is essential to limit the errors associated to the operations switches. In general, these can be minimized by using the smallest geometries allowed by the technology. In the context of a CS-based acquisition, the problems are accentuated by the fact that the array capacitor has to be divided in n_b slices. Being smaller than the entire array, each element is more sensitive to noise. Moreover, the modulated samples have to be preserved for an entire window, therefore leakage currents are of great concern.

4.1. Errors on commutations

The two errors typically associated to the commutations of the switches are charge injection and clock feedthrough.

In order for a MOS transistor to change its conduction state, the channel has to be created/destroyed, implying a transfer of charge with the surrounding elements. When the transistor is driving a hold capacitor, the exchange leads to a voltage drop on the capacitor whose amount depends on the charge and the size of the capacitor itself. Since the charge stored inside the channel is a function of the local potential, modeling the injection effect is challenging as it would require a pointwise, time-dependent description of the channel potential as the gate voltage is modified [10].

If the transition of the gate voltage is fast enough, a reasonable approximation is to consider the charge equally divided between source and drain (Figure 4.1).

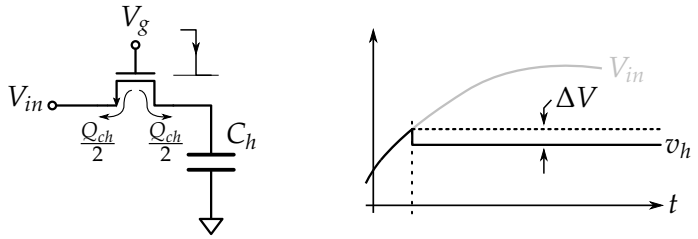

Figure 4.1.: Charge injection and its effects on the sampled value

The voltage drop induced in the hold capacitor C_h is

$$\Delta V = \frac{WLC'_{ox}}{2C_h}|V_{g,on} - V_{in}|.$$

Here W and L are the transistor dimensions and C'_{ox} the oxide capacitance per unit area. Being input dependent, it leads to a distortion of the reconstructed waveform.

Other than increasing the value of the hold capacitor and minimizing the channel area, a couple of techniques can be employed to limit the errors. The first and most effective one involves the use of dummy switches (Figure 4.2a). The dummy element is driven in phase opposition with respect to the main switch, so that when the inversion layer is removed from the main switch, the charge is absorbed entirely by the dummy.

The technique is very effective only if the clock transitions are fast enough, so that the half-splitting approximation holds. Thus the dummy size has to be half that of the main transistor. It is important to ensure that the injected charge is captured by the dummy element, therefore a small delay in the driving signal of the dummy has to be introduced so that the inversion layer in it is formed after the main switch is opened.

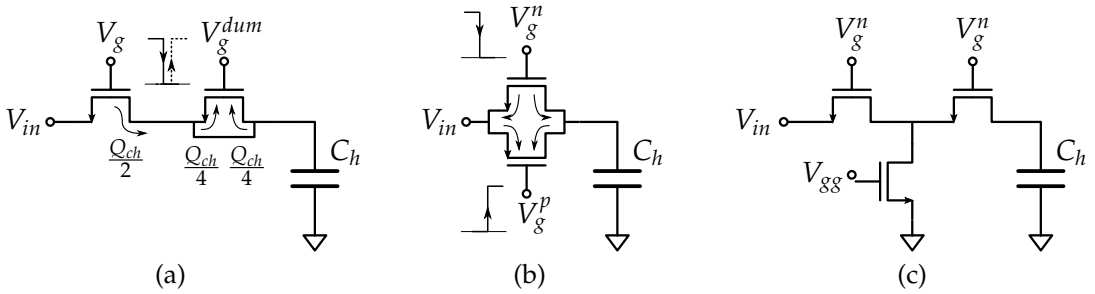

Figure 4.2.: (a-b) Charge injection compensation techniques: (a) dummy transistor and (b) transmission gate. (c) T-switch configuration

A second solution is the use of a transmission gate (Figure 4.2b). This is typically employed when the input signal varies in the entire supply range. Since transistor of opposite polarity have inversion charges of opposite kind, when both of them are turned off, they inject charges that mutually compensate. However the solution is not as effective as the first one, since the channel charge depends on V_{gs} and the two transistors are turned on by opposite voltages.

The other effect induced by commutations of the switches is due to the capacitive coupling from the gate terminal to the hold capacitance. A capacitive partition takes effect, so that a voltage variation is observed on the hold capacitor. Also in this case, minimum size switches minimize the problem. The two techniques already seen, requiring control voltages in phase opposition, also lead to a reduction of the feedthrough effect. An additional capacitive coupling to be considered is from source to drain. Fast transitions of the input may be felt by the hold capacitor even when the switches are open. A structure that reduces the effect to a minimum is the so called T switch Fig. 4.2c. In it, the pass transistors are doubled and the intermediate node, when both the switches are open, is grounded. The low impedance of the node shields the hold capacitor from the noise injected by the input signal.

4.2. Considerations on leakage currents

The mere presence of the switches leads to a continuous discharge of the capacitors because of subthreshold conduction as well as the reverse current through the source/drain diffusions. Minimizing the junction area is the first step to reduce the loss. However, the hold time in this CS-based applications is a significant fraction of one ECG period, i.e. 1 s. In Section 2.3.1 we have already showed how reconstruction is impossible operating on a time scale this large. As an additional remark, the solutions proposed here to minimize the noise injected by the switches are detrimental from the point of view of leakage, since the number of junctions seen by the hold capacitors increases. The solution using the dummy element is the worse, with four junctions of the same kind acting on the hold node.

5. Compensator

Having established the origins of leakage (Section 2.3.1) and its effect on CS reconstruction (Section 4.2), it is clear that in order for the AIC to actually work, the architecture has to perform some kind of compensation of the leakage currents. A few solutions have been published in the literature e.g. [12, 13]. Among them, the one in [14] has been chosen because it is robust, compact and power efficient. The topology is shown in Figure 5.1.

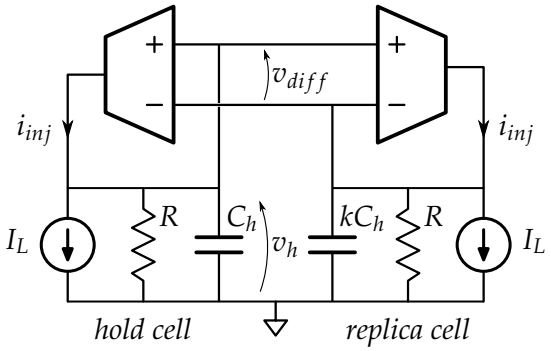

Figure 5.1.: Topology of the compensator. The left branch is the original hold cell, the other is the scaled-down replica

A hold cell is modeled as a capacitor C_h in parallel to the equivalent off resistance of the switches R and a source of leakage I_L. A downscaled replica kC_h of the hold cell C_h is characterized by a smaller value of capacitance, being $k < 1$, and identical leakage currents and off resistance, since switches should be identical and placed in close proximity on the silicon die. Having equal currents in unequal capacitances, the voltage rate of change will be higher in the scaled replica. Starting from the same initial voltage, i.e. the one sampled from the input, a difference will grow across the cells.

This difference is converted by two identical transconductors and injected equally back in the cells, reducing the rate of change. Intuitively (neglecting the effects of R), when the voltage difference results in a compensation current exactly equal to the leakage one, the variation stops and the voltage can be held indefinitely. If the compensation is insufficient, the voltage difference grows in magnitude, increasing the injected current. Conversely, if injection is too much, both voltages increase, at a faster rate on the smallest capacitor. Feedback then reduces the injection. In any case, the effect is to bring the compensation current towards the level of the leakage. In reality,

the always-present resistive component prevents the settlement to an equilibrium and both voltages decay continuously over time. Tuning the parameters of the compensator, namely the scaling factor k and transconductance g_m, or even modifying the value of C_h, the voltage variation can be constrained to an acceptable level.

The complete analytic model of the compensator will be derived from its block diagram description in Laplace domain. This will allow the evaluation of the time domain behaviour as well as stability margins. Doing so considering the effect of initial conditions requires the description of a charged capacitor in Laplace domain. The development is included here in order to select the most suitable form of its description.

5.1. Charged Capacitor in Laplace Domain

Starting from the constitutive relation of a linear capacitor

$$i(t) = C\frac{dv(t)}{dt},$$

we get:

$$\mathscr{L}\{i(t)\} = C\left(s\mathscr{L}\{v(t)\} - v(0^-)\right)$$

$$I(s) + Cv(0^-) = sCV(s) \tag{5.1}$$

$$I(s) = sC\left(V(s) - \frac{v(0^-)}{s}\right) \tag{5.2}$$

According to (5.1), initial conditions can be represented as an independent current source acting in parallel to the capacitor, since the term $Cv(0^-)$ does not depend on quantities in the transformed domain. The equivalent circuit and its Thévenin counterpart (5.2) are shown in Figure 5.2.

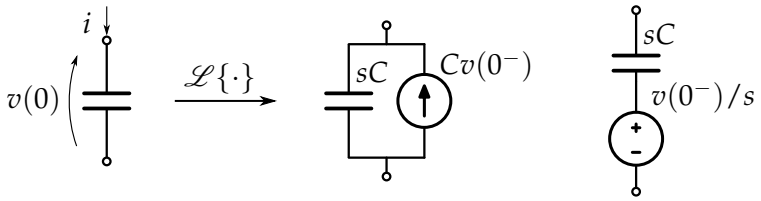

Figure 5.2.: Equivalent circuits of a charged capacitor in Laplace domain.

Notice how the equivalent current in (5.1) has units of charge (capacitance times voltage). This is indeed the case since the Laplace transform integrates a physical quantity, here the capacitor current, over time.

Between the two possible descriptions of a charged capacitor, the one with the parallel current source is easier to apply to the actual structure of the compensator, as each cell is already made of parallel elements and the addition is straightforward.

5.2. Block diagram description

The most general description of the compensator has to account for all possible inputs, including asymmetries between the two cells. The system then becomes a Multiple-Input-Single-Output one, since the only quantity of interest is the hold voltage $V_h(s)$. As such, the compensator is described by one transfer function for every input.

The complete block diagram is shown in Figure 5.3. The nominal leakage $I_L(s)$ is common to both cells. An eventual difference $\Delta I_L(s)$ may unbalance one of the branches. The initial voltage has an effect which depends on the capacitance (5.1), thus the different scaling of that input before being injected into the cells. A second asymmetry is that of the sampled voltage, represented by ΔV_0.

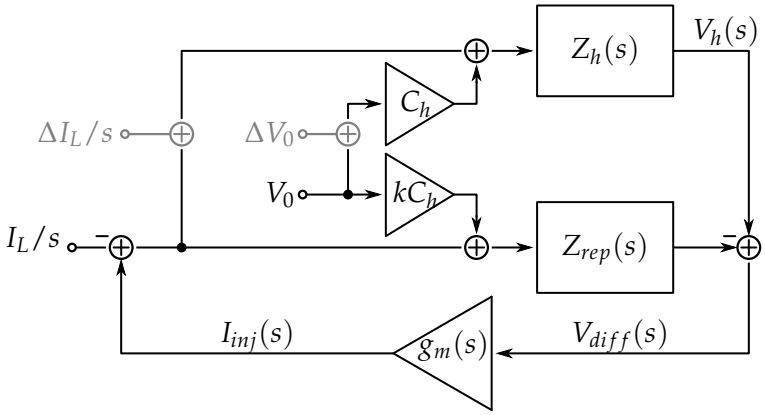

Figure 5.3.: Complete block diagram of the compensator

The hold and replica cells are represented by their equivalent impedance, given by the parallel connection of the capacitor and the resistor. Therefore:

$$Z_h(s) = \frac{1}{sC_h} \oplus R_h = \frac{R_h}{1 + sR_hC_h}$$
$$Z_{rep}(s) = \frac{1}{skC_h} \oplus R_{rep} = \frac{R_{rep}}{1 + skR_{rep}C_h}$$

Concerning the description of input signals, initial conditions in time domain are described by a Dirac delta function, which translates, coherently with the result in (5.1),

to a constant term in Laplace domain:

$$V_0(s) = V_0$$
$$\Delta V_0(s) = \Delta V_0.$$

Leakage instead can be considered a step input. Until the switches are closed and the hold capacitor tracks the input signal, there is a low impedance path for the leakage currents, which flow preferentially towards the signal source. As soon as the switches are opened, the current flows entirely in the capacitive cells. Therefore:

$$I_L(s) = \frac{I_L}{s}.$$

The same is true for the asymmetry $\Delta I_L(s)$.

Having defined all the elements in Figure 5.3, the input/output transfer functions can be derived (the complex frequency variable has been hidden):

$$\begin{aligned} H|_{V_0} &= Z_h C \frac{1 + g_m Z_{rep}\,(1-k)}{1 + g_m \left(Z_{rep} - Z_h\right)} \\ H|_{I_l} &= \frac{-Z_h}{1 + g_m \left(Z_{rep} - Z_h\right)} \\ H|_{\Delta I_l} &= \frac{g_m Z_h Z_{rep}}{1 + g_m \left(Z_{rep} - Z_h\right)} \\ H|_{V_{off}} &= \frac{g_m Z_h}{1 + g_m \left(Z_{rep} - Z_h\right)} \end{aligned} \tag{5.3}$$

Depending on the accuracy of the description one wants to obtain, several levels of approximation can be derived by considering particular values for the system parameters.

In the following analysis, only the first two transfer functions will be considered, deriving the expression of the hold voltage in time domain in the case of constant transconductance $g_{m,0}$.

5.2.1. Infinite parallel resistances

Having both R_h and $R_{rep} \to \infty$, the previous expressions become:

$$\begin{aligned} H|_{V_0} &\to \frac{1}{s} \\ H|_{I_l} &\to -\frac{k}{g_{m,0}\,(1-k)} \frac{1}{1 + s\frac{Ck}{g_{m,0}(1-k)}} \end{aligned}$$

and the corresponding components of $v_h(t)$ are:

$$\begin{aligned} v_h(t)|_{v_0} &= \mathscr{L}^{-1}\left\{v(0^-)\frac{1}{s}\right\} \\ &= v(0^-)u(t) \\ v_h(t)|_{i_l} &= \mathscr{L}^{-1}\left\{-\frac{I_L}{g_m\frac{1-k}{k}}\frac{1}{s}\frac{1}{1+s\frac{Ck}{g_{m,0}(1-k)}}\right\} \\ &= -\frac{I_L}{g_{m,0}\frac{1-k}{k}}\left[1-\exp\left(-\frac{t}{\frac{Ck}{g_{m,0}(1-k)}}\right)\right]u(t) \end{aligned}$$

Obviously, since the leakage current stems from reverse biased junctions, the discharge would end once the voltage across the junctions is null. The model is thus valid until the overall hold voltage, summing all contributions, is $v_h(t) = 0$.

Notice that the initial conditions are maintained indefinitely. Conversely, leakage results in a sudden but small discharge, up to a voltage difference between the two cells that compensates completely the discharge current. Therefore the voltage waveform, shown in Figure 5.4, after an initial transient (unnoticeable at the scale of the plot), remains constant. Compare it to the uncompensated cell which discharges completely in a much shorter time.

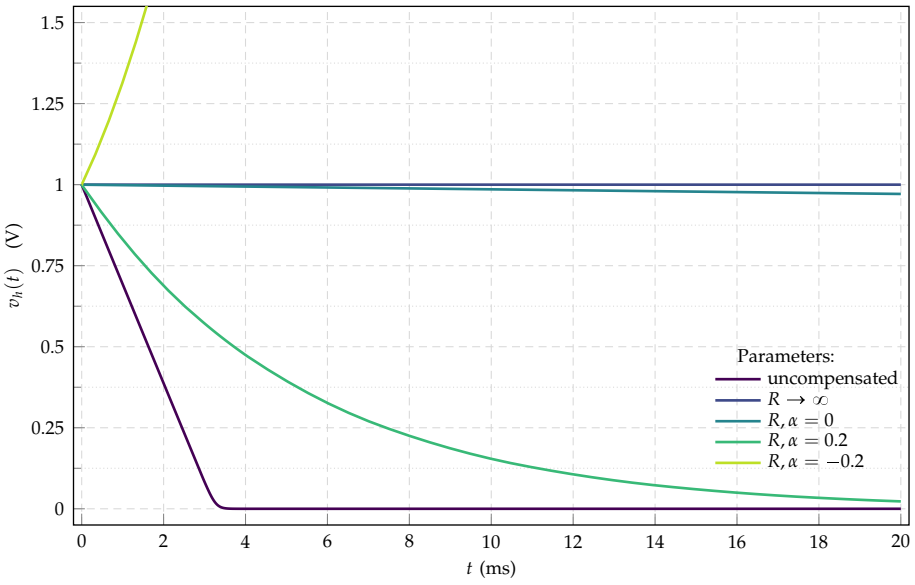

Figure 5.4.: Transient behaviour of the voltage in the hold cell. $R = 1\,\mathrm{G\Omega}$ when finite, $g_{m0} = 1\,\mu\mathrm{S}$, $C_h = 1\,\mathrm{pF}$, $k = 0.1$ and $I_L = 300\,\mathrm{nA}$.

The magnitude and phase plots of the transfer function $H|_{I_l}(s)$ are shown in Fig. 5.5.

5.2.2. Finite and equal parallel resistances

With identical resistances $R_h = R_{rep} = R$ and having $g_{m0}R \gg 1$, the transfer functions become:

$$H|_{V_0} = g_{m,0}R\,(1-k)\,RC\frac{1}{1+sg_{m,0}R\,(1-k)\,RC}$$
$$H|_{I_l} = -R\frac{1+skRC}{1+sg_{m,0}R\,(1-k)\,RC}$$

The time domain behaviour is:

$$v_h(t)|_{v_0} = V_0 \exp\left(-\frac{t}{g_{m,0}R\,(1-k)\,RC}\right)u(t)$$
$$v_h(t)|_{i_l} = -I_L R\left[1-\exp\left(-\frac{t}{g_{m,0}R\,(1-k)\,RC}\right)\right]u(t) \tag{5.4}$$

Adding a finite resistance, as it is reasonable to do, results in a larger decay of the voltage. The interesting result to highlight is that the compensator determines a scaling of the time constant by a factor $g_{m,0}R(1-k)$. Knowing the duration of the interval in which a sample has to be maintained, these parameters can be sized to minimize the decay.

5.2.3. Mismatched parallel resistances

Under the effect of an asymmetry α in the equivalent parallel resistances, the impedances of the cells become:

$$Z_H(s) = \frac{R}{1+sRC}$$
$$Z_{REP}(s) = \frac{R\,(1+\alpha)}{1+sRC\,(1+\alpha)}$$

Correspondingly, the transfer functions are:

$$H|_{V_0} = g_{m,0}R\,(1-k)\,RC\frac{1}{1+sg_{m,0}R\,(1-k)\,RC}$$
$$H|_{I_l} = -\frac{R}{1+g_{m,0}R\alpha}\frac{1+skRC\,(1+\alpha)}{\left(1+s\frac{g_{m,0}R(1+\alpha-k)}{1+g_{m,0}R\alpha}RC\right)\left(1+s\frac{C(1+\alpha)k}{g_m(1+\alpha-k)}\right)}$$

In time domain this corresponds to:

$$v_h(t)|_{v_0} \simeq V_0 \exp\left(-\frac{t}{g_{m,0}R\,(1-k)\,RC}\right) u(t)$$

$$v_h(t)|_{i_l} \simeq -\frac{I_L R}{1+g_{m,0}R\alpha}\left[1-\exp\left(-\frac{t}{\frac{g_{m,0}R(1+\alpha-k)}{1+g_{m,0}R\alpha}RC}\right)\right] u(t)$$

$$-I_L\frac{k}{g_{m,0}RC}\left[1-\exp\left(-\frac{t}{\frac{C(1+\alpha)k}{g_m(1+\alpha-k)}}\right)\right] u(t)$$

The decay of initial conditions is not affected significantly if α is small, as it is a decay of two RC cells of almost equal component values.

If the term $g_{m0}R\alpha$ is larger than 1 in absolute value, the mismatch factor reduces the time constant of the decay. The important point is that, as α can also be negative, the exponential decay may become an exponential growth, with the consequence of a much larger absolute variation of the signal in a given time interval.

The behaviour is identical at high frequencies, whereas at low frequencies the nature of the resistances heavily affects the curves, especially in terms of phase. In any case, only real poles have emerged, therefore no oscillations can be observed on the output voltage.

5.3. CS reconstruction with compensated leakage

All the discussion so far has been focused on the behaviour of the compensator, trying to model analytically all the effects we could possibly expect from such a feedback circuit. What we are really interested in is how the reconstruction of a signal acquired by the AIC compares with the original one.

As already done in Section 2.3.1 to evaluate the global effects of leakage, the time domain model defined in(5.4) has been applied to the sensing matrix S. Using as parameters $I_L = 300$ pA, $R = 1\,\mathrm{G}\Omega$, $g_{m0} = 1$ mS, $k = 0.1$ and $C = 1$ pF, the dashed curves in 5.6 have been obtained. The plot includes the curves already shown in the previous figures so that it is easier to draw a comparison.

The first thing to notice is that the lighter curves are quite close to the original, solid lines. Compensation results in a good performance if the width of the matrix blocks is not too long, i.e. the acquisition window does not span too much time. Indeed for $n_b = 8$ the compensated and reference curves are identical. For CR > 2.5 also the $n_b = 16$ curves coincide and result in better performance with respect to the previous ones. Increasing n_b further, compensation starts to suffer from low gain, especially since the total capacitance of the converter has been kept constant. Only for CR > 3.5 the

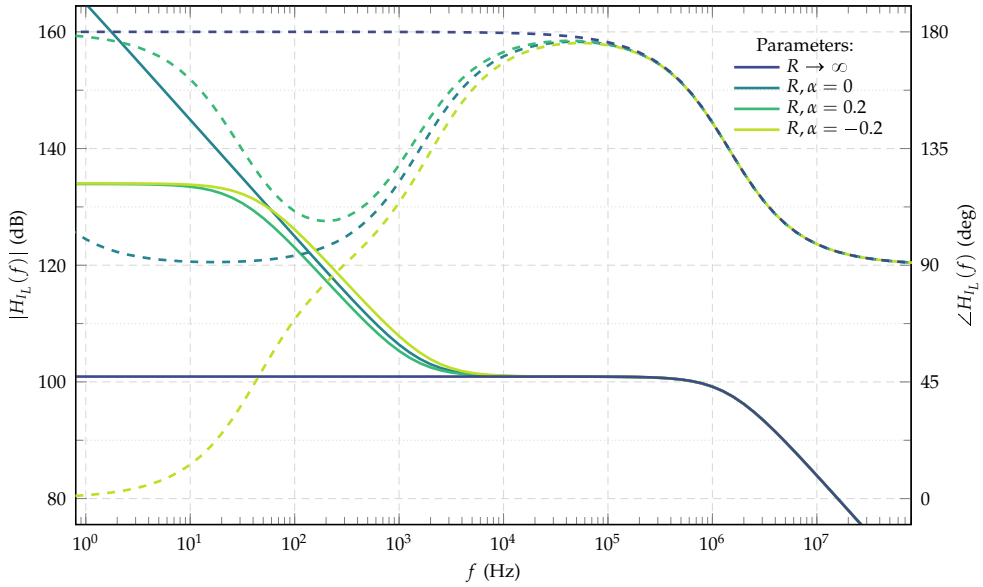

Figure 5.5.: Transfer function from input I_L for several values of the parameters (specified in the legend). Dashed curves refer to the phase. $R = 1\,\mathrm{G\Omega}$ when finite, $g_{m0} = 1\,\mu\mathrm{S}$, $C_h = 1\,\mathrm{pF}$, $k = 0.1$.

quality improves above the previously discussed ones. The best compromise between hardware complexity and reconstruction quality, for low compression ratios, is $n_b = 16$.

It is striking to observe how the ARSNR is still negative in the case of a full sensing matrix. Such a long acquisition window (approx one second), coupled with extremely small hold capacitances, would require a much better, probably unfeasible, compensator. The result is that all the information is lost below the noise level.

5.4. Stability

One simplification that in reality does not hold is the constant transconductance, although it has been helpful to observe and understand the fundamental nature of the compensation. The frequency response of the transconductor may bring the system at the onset of instability, giving rise to unwanted oscillations. Instead of following the same approach as before, deriving the time-domain behaviour of the compensator through the transfer functions, we will take advantage of the results from the theory of feedback systems [15], observing the properties of the loop gain $T(s)$.

$T(s)$ is the gain incurred by a signal traveling around the loop. It can be evaluated from the block diagram in 5.3 by cutting the loop at one point, injecting a test signal on the forward path and evaluating the response coming back at the cut. However,

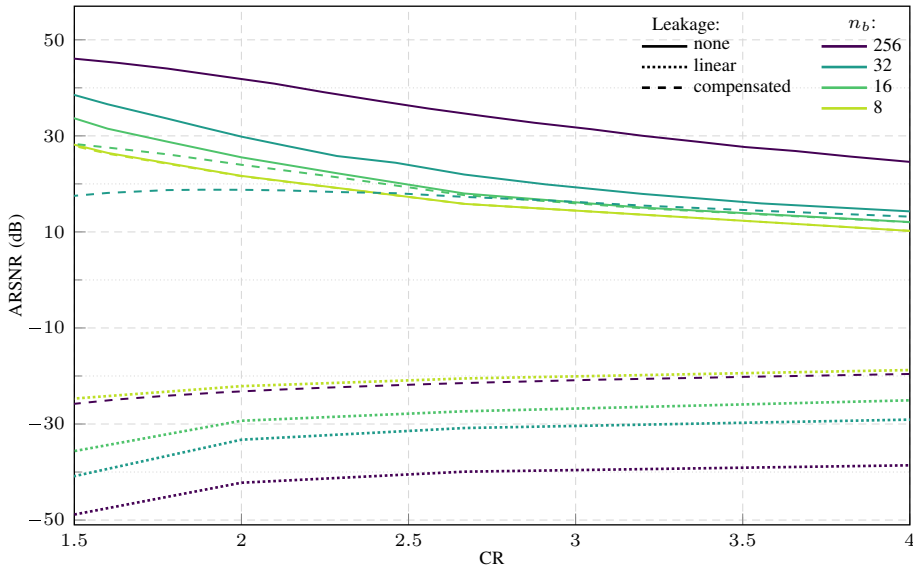

Figure 5.6.: Reconstruction performance with leakage compensation

changing the location of the injection, the loop gain may change. Thus one might ask if different gains result in the same stability criteria. Indeed this is case, as derived later in 5.4.1.

Even if the loop gain changes, the transfer functions have to be the same, since the system has not been modified. According to [16] the complete expression of a transfer function requires, other than T, also the two quantities H_∞ and H_0, which are affected by the choice of injection point in a way that keeps the overall transfer function unchanged.

Having already derived the transfer functions in a previous section, we can readily observe one form of the loop gain. The denominators of (5.3) are in fact equal to $1 + T(s)$. That form of the loop gain is the one obtained by injecting a test signal as a current just after the transconductor, as confirmed by a direct visual inspection of the block diagram. Therefore:

$$T(s) = g_m(s)\left(Z_H(s) - Z_{REP}(s)\right).$$

We will use this form as it is the easiest to work with and understand (i.e. it has "low-entropy" [17]). The behaviour of T as frequency is increased is shown in Figure 5.7, still in the case of a constant transconductance.

When the parallel resistance goes to infinity, the system behaves as an ideal integrator, therefore even a minuscule amount of current injected in the capacitors gives rise to a large voltage, easily detectable by the transconductor that can compensate for it.

Having finite and equal resistances, the DC component of the current does not flow in the capacitors anymore. However, identical resistances affected by the same current generate equal voltages, making the loop insensitive to the effect, thus the zero in the

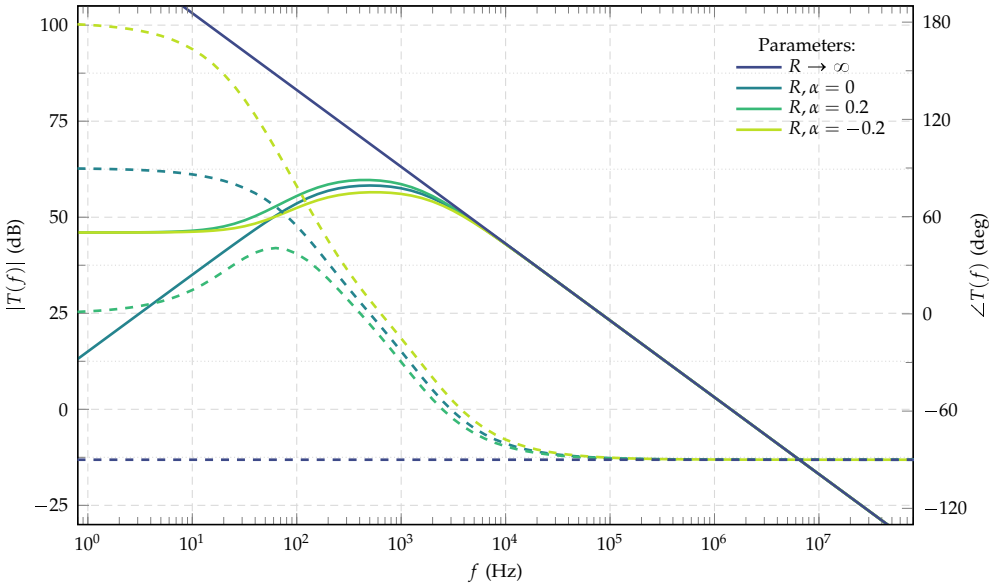

Figure 5.7.: Loop gain with constant transconductance. Dashed curves refer to the phase.

origin.

Introducing a mismatch in the resistive element, the DC gain becomes finite. The phase depends on the sign of the mismatch. If negative, there is half a rotation of the signal around the loop which, coupled with the sign inversion due to the negative feedback, enhances the voltage variation. This is what has been observed in Fig. 5.4 with the exponential growth of the sampled voltage.

Let us now model the transconductor with a transfer function of the kind:

$$g_m(s) = g_{m,0} \frac{1 + \frac{s}{\omega_z}}{\left(1 + \frac{s}{\omega_{p,1}}\right)\left(1 + \frac{s}{\omega_{p,2}}\right)} \tag{5.5}$$

The most critical condition is for the transfer function to have a right half plane zero, followed by two poles with negative real part, especially in combination with a small scaling factor k for the capacitances. All the singularities contribute a decrease in phase, which may reach $-180°$ when the gain is still larger than 1, $|T| > 0\,\text{dB}$. In that condition, the system may start to oscillate.

The same might be concluded by observing the sensitivity function, defined as:

$$S(s) \stackrel{\text{def}}{=} \frac{1}{1 + T(s)} \simeq \begin{cases} \frac{1}{T(s)} & \text{if } |T(s)| \gg 1 \\ 1 & \text{if } |T(s)| \ll 1 \end{cases} \tag{5.6}$$

$S(s)$ can be used to easily obtain a closed loop transfer function from the corresponding open loop one, evaluated on the direct path from the input to the output under

consideration:

$$H_{cl}(s) = S(s)H_{ol}(s)$$

If H_{ol} is well behaved, i.e. it does not present unstable poles, S, which stems from the introduction of a feedback loop, is the one that may bring unwanted phenomena in play. Indeed, the transition between the two approximations in (5.6) is the critical one. Since typically T decreases with frequency, we can safely assume that $S \simeq 1$ for frequencies above a certain crossover value. Depending on the slope with which T approaches such a value, a given number of singularities have to appear to result in a flat behaviour at high frequencies.

If the slope of T at the 0 dB crossing point is -20 dB/dec, then the sensitivity function will have only one pole at that frequency. In the case of a larger slope, two or more poles will appear, with the possibility of them being complex conjugate. This will be reflected in the transient response showing peaking and oscillations of $v_h(t)$.

5.4.1. Equivalence of any loop gain

It is known [16] that a given loop may have different loop gains depending on the injection point from which the gain is evaluated. By using the complete expression of the gain established by Prof. Middlebrook, it is clear that also the terms G_∞ and G_0 have to change in order for the transfer function to be the same. Expanding each quantity in the expression of G in terms of its numerator and denominator, the following result is obtained

$$\begin{aligned} G = \frac{G_\infty T + G_0}{1+T} &= \frac{\frac{N_\infty}{D_\infty}\frac{N_T}{D_T} + \frac{N_0}{D_0}}{1 + \frac{N_T}{D_T}} \\ &= \frac{N_\infty N_T D_0 + D_\infty D_T N_0}{D_\infty D_0 (D_T + N_T)} \end{aligned}$$

Poles of G_∞ and G_0 are also poles of the transfer function, therefore stable open loop systems still have stable poles coming from those terms. The remaining part of the denominator is the one typically analyzed through the Nyquist criterion. Because the stability of certain G is unique, even if unknown, and provided that the two gains do not have unstable singularities, the conditions described by any T have to agree.

6. Comparator

The element responsible for the evolution of the successive approximation conversion is the comparator. As for any other circuit, the energy it requires should be minimal, especially in the context of a CS-based acquisition, where one of the key factors motivating its introduction is that of saving power. An additional parameter is the response time from the stimulation of the comparator until a decision is made, since it determines the speed at which bits can be generated. Even more importantly, the internal coupling from the comparator outputs to its inputs has to avoid the injection of noise back in the hold capacitors. And finally the comparator should be able to resolve an input voltage smaller than the expected quantization error of the D/A converter.

All these constraints drove the selection of a suitable topology. The candidate architecture we have analyzed in detail and that will be described in the following is shown in Figure 6.1 and has been taken from [18]. It is built as the cascade of a dynamic residual preamplifier and a parallel-coupled regenerative latch.

Each stage will be analyzed separately, deriving the analytic expression of the transient response in the different operating phases. Then, asymmetries of the circuits will be quantified as an input referred offset voltage.

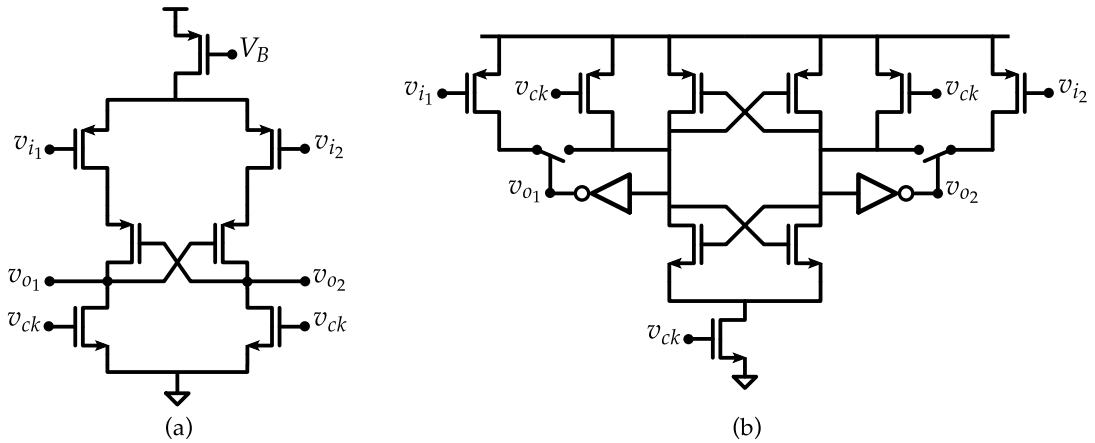

Figure 6.1.: Elements of the comparator: (a) dynamic residual preamplifier and (b) parallel-coupled regenerative latch.

6.1. Preamplifier

The first stage of the comparator is a dynamic residual amplifier. Its main purpose is to decouple the fast transitions of the comparator outputs from the highly sensitive inputs. It works by unbalancing the current flowing in two identical, capacitively loaded branches. The output voltage difference, observed on the capacitors, grows over time, resulting in a time dependent voltage gain. The cross-coupled transistors in the middle of each branch introduce some positive feedback, increasing the gain until one branch saturates.

The circuit shown in Figure 6.2a, requires steady inputs, which the capacitive DAC is able to provide. They are applied to the differential couple M_1, M_2, which charges the parasitic capacitance C_l of two clocked MOS transistors M_5, M_6. When the clock signal is active, the output nodes are shorted to ground, removing any memory of the previous cycle. Releasing the clock, the capacitances between drain and ground become the load for each branch (Fig. 6.2).

The cross-coupled pair (M_3, M_4), is initially current-biased and contributes a little gain. Its most significant effect is the positive feedback introduced during the regenerative phase, turning the weakest branch off and resulting in further amplification.

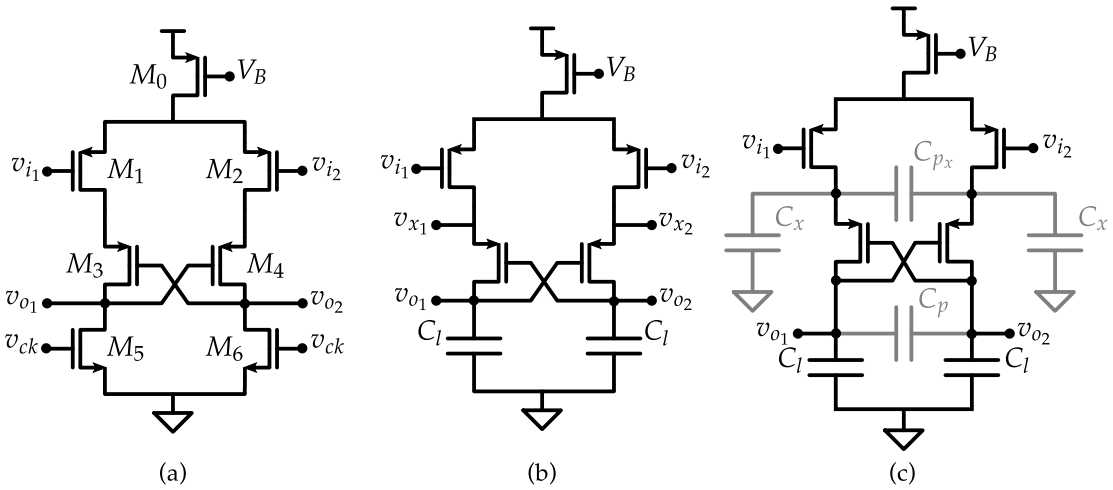

Figure 6.2.: (a) Preamplifier circuit. (b) Equivalent circuit in the active phase. (c) Equivalent circuit with parasitic capacitances.

Assumptions in the analysis are:

- quadratic MOSFET model, neglecting channel length modulation
- exact symmetry of the circuit (mismatches will be considered separately)
- small input differential voltage, as this is the most critical operating condition for the amplifier

- parasitic capacitive loading at the inner nodes $X_{\{1,2\}}$, as well as across the branches

Under these assumptions, the circuit can be linearized around its DC operating point, obtaining Common-Mode (CM) and Differential-Mode (DM) equivalent circuits. Given a couple of signals y_1 and y_2, they can be described in an equivalent form as:

$$\begin{aligned} y_1 &= y_{cm} + \frac{y_{dm}}{2} \\ y_2 &= y_{cm} - \frac{y_{dm}}{2}, \end{aligned} \qquad \text{where} \qquad \begin{aligned} y_{cm} &= \frac{y_1 + y_2}{2} \\ y_{dm} &= y_2 - y_1. \end{aligned}$$

The analysis is simplified if the branches are decoupled, removing the crossed connection involving M_3 and M_4 in such a way as to preserve the behaviour of the original circuit.

In the initial phase of linear output voltage growth, assuming all devices in saturation, the crossed couple acts as a voltage shifter from the output on one branch, to the X node on the opposite branch. Therefore it can be represented as an equivalent voltage source V_{sh} acting on a single branch, with a proper value. Let us first write down the voltage at nodes X:

$$\begin{aligned} v_{x_1} &= v_{o_2} + V_{thp} + \sqrt{\frac{2i_1}{\beta_{xcp}}} \\ v_{x_2} &= v_{o_1} + V_{thp} + \sqrt{\frac{2i_2}{\beta_{xcp}}} \end{aligned}$$

We want to express it in terms of the output voltage on the same branch:

$$\begin{aligned} v_{x_1} &= v_{o_1} + V_{sh_1} \\ v_{x_2} &= v_{o_2} + V_{sh_2} \end{aligned}$$

Solving for V_{sh} and linearizing the square root of the current, considering a small differential voltage v_d applied to the input of the amplifier, we obtain

$$\begin{aligned} V_{sh_1} &= v_{o_2} - v_{o_1} + V_{\text{thp}} + \sqrt{\frac{2i_1}{\beta_{\text{xcp}}}} \\ &= v_o^{\text{dm}} + V_{\text{thp}} + \sqrt{\frac{I_B}{\beta_{\text{xcp}}}} \left(1 + \frac{g_m^{\text{diff}} v_d}{I_B}\right) \\ &= V_{\text{thp}} + \sqrt{\frac{I_B}{\beta_{\text{xcp}}}} + v_o^{\text{dm}} + \sqrt{\frac{\beta_{\text{diff}}}{\beta_{\text{xcp}}}} v_d \\ V_{sh_2} &= V_{\text{thp}} + \sqrt{\frac{I_B}{\beta_{\text{xcp}}}} - v_o^{\text{dm}} - \sqrt{\frac{\beta_{\text{diff}}}{\beta_{\text{xcp}}}} v_d \end{aligned} \tag{6.1}$$

In these expressions, V_{th} is the threshold voltage of the devices, I_B the bias current determined by M_0 and β the transconductance of the MOS devices.

The first two terms in the result belong to the common mode, the last two are instead differential terms. Since the expressions are mutually-decoupled, the CM-DM half circuits can be easily obtained (Figure 6.3). In the first, parasitic 'across' capacitances C_p and C_{p_x} see the same voltage on both terminals, therefore they can be safely removed. In the latter, those same capacitances see the positive differential voltage on one side, and the negative one on the other, therefore, using a single branch, the equivalent capacitive loading has to double.

6.1.1. Linear growth regime

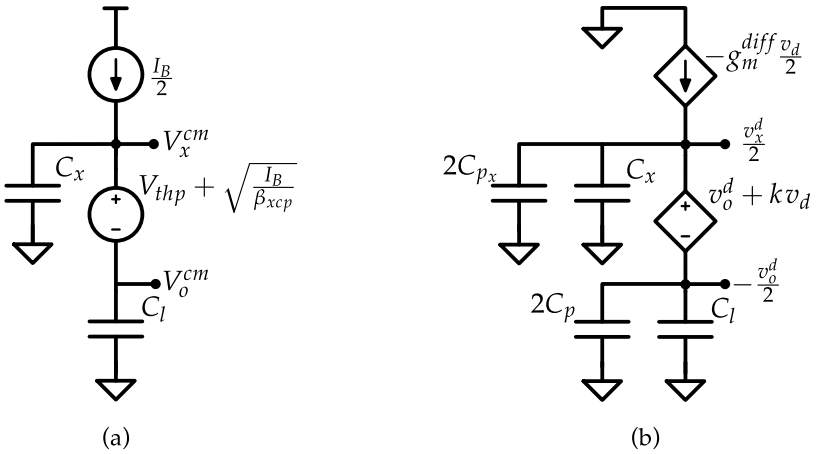

Figure 6.3.: Preamplifier equivalent half-circuits during the linear growth phase. (a) CM circuit (b) DM circuit.

Common mode:

Applying only the common mode input voltage, the differential couple injects half of the bias current in each branch. The crossed couple introduces a voltage shift of value $V_{\text{thp}} + \sqrt{I_B/\beta_{\text{xcp}}}$ between the output and the inner node. The output voltage is then

$$V_o^{\text{cm}}(t) \simeq \frac{I_B}{2(C_x + C_l)}t.$$

The linear increase in voltage goes on until a time t_1, when the differential couple leaves saturation. Assuming $v_o^{\text{dm}}(t_1)$ still small compared to V_o^{cm}, the time t_1 can be approximated by imposing the condition $V_x^{\text{cm}}(t_1) = V_{\text{in}}^{\text{cm}} + V_{\text{thp}}$, obtaining

$$t_1 \simeq \frac{V_{\text{in}}^{\text{cm}}}{I_B} 2(C_x + C_l),$$

where the overdrive voltage of the crossed couple has been neglected and the threshold voltages of the crossed and differential couples have been considered equal. Under the same assumption of small $v_o^{\mathrm{dm}}(t_1)$, then both branches of the differential couple have similar voltages, therefore they leave saturation almost at the same time.

Differential mode:

In differential mode, the input transistors are ground-referred transconductors, being their source terminal equivalent to virtual ground (Figure 6.3b). The crossed couple, according to (6.1), is modeled as a voltage-controlled voltage source, depending on both the input and output differential voltages. This substitution unveils the origin of the positive feedback. While the output changes as $-v_o^{\mathrm{dm}}/2$, the inner node goes in the opposite direction, being $+v_o^{\mathrm{dm}}/2 + kv_d$. If v_{o_1} is the lowest output voltage, v_{x_1} is the highest inner voltage, weakening the differential couple on its branch and further decreasing the growth of v_{o_1}. The differential output voltage is then

$$v_o^{\mathrm{dm}}(t) = \frac{g_m v_d}{C_l - C_x + 2(C_p - C_{p_x})} t$$

At time t_1, when the linear growth ends, its value is

$$\begin{aligned} v_o^{\mathrm{dm}}(t_1) &\simeq g_m v_d \frac{V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} \frac{2(C_l + C_x)}{C_l - C_x} \\ &\simeq 2 g_m v_d \frac{V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} \end{aligned}$$

where last approximation holds if C_x is small compared to C_l.

6.1.2. Regenerative regime

Let us reconsider the original circuit (Fig. 6.2a) with a trioded differential couple. The effects of v_d can be neglected and the couple can be modeled as two equal-valued resistors R acting as source degeneration for the crossed pair. The differential voltage across the internal nodes is then

$$\begin{aligned} v_x^{\mathrm{dm}} &= v_o^{\mathrm{dm}} + \sqrt{\frac{I_b}{\beta_{\mathrm{xcp}}}} \left(-\frac{i_1 - i_2}{I_B} \right) \\ &= v_o^{\mathrm{dm}} - \sqrt{\frac{I_b}{\beta_{\mathrm{xcp}}}} \frac{v_x^{\mathrm{dm}}}{R I_B} \\ &= \frac{v_o^{\mathrm{dm}}}{1 + \sqrt{\frac{I_b}{\beta_{\mathrm{xcp}}}} \frac{v_x^{\mathrm{dm}}}{R I_B}} \\ &\simeq \frac{g_m^{\mathrm{xcp}} R}{1 + g_m^{\mathrm{xcp}} R} v_o^{\mathrm{dm}} \end{aligned}$$

Being the resistance in the order of 10 kΩ, and the transconductance a few to tens of µS, v_x^{dm} is much smaller than v_o^{dm}. Neglecting it, the sources of the crossed couple are both at virtual ground. The equivalent circuit then is the one in Figure 6.4a.

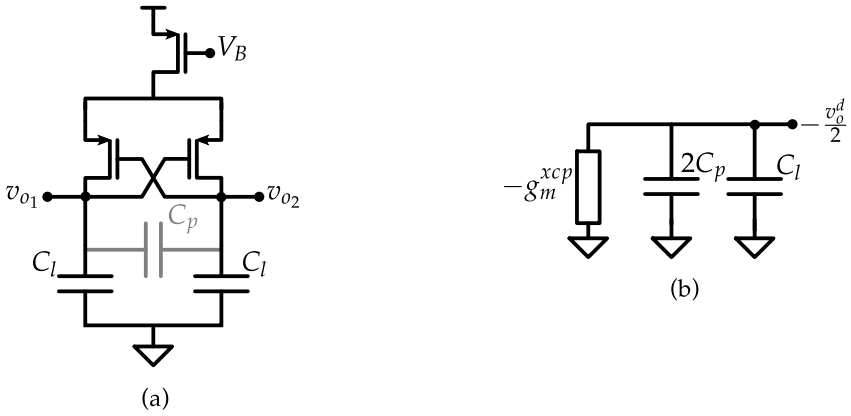

Figure 6.4.: Preamplifier in regeneration

Transistors M_3 and M_4 act as transconductors, driven by the voltage on the opposite branch. Since the output voltages at the end of linear growth are slightly unbalanced, the currents through the branches, depending on the conduction of the cross-coupled transistors, are unbalanced, too. For simplicity, let us assume the transconductance to be constant.

Since the control voltage of each transconductor is the output on the other branch, the transistors behave as a negative conductance, as shown in Figure 6.4b. The time constant of the circuit is negative, resulting in an exponential growth of the differential output voltage:

$$v_o^{\mathrm{dm}}(t) = v_o^{\mathrm{dm}}(0) \exp\left(\frac{g_m^{\mathrm{xcp}}}{C_l + 2C_p} t\right)$$

Being the total capacitance relatively small, the transient immediately determines a voltage sufficient to turn one branch off (by reducing the source/drain voltage) and bring the transistor on the other side into triode region. Such a condition can be expressed as $v_o^{\mathrm{dm}}(t) = V_{\mathrm{thp}}$ and it is reached after

$$\Delta t = \frac{C_l + 2C_p}{g_m^{\mathrm{xcp}}} \ln \frac{V_{\mathrm{thp}}}{v_o^{\mathrm{dm}}(t_1)}$$

In reality the transconductance decreases rapidly during the transient, causing the transient to last longer than predicted.

The fundamental result is that even a small v_o^{dm} is rapidly amplified to a value close to one threshold voltage, large enough to make the second stage of the comparator insensitive to mismatches. To achieve that, we have to guarantee the correct sign of

$v_o^{\mathrm{dm}}(t_1)$ with respect to v_d, thus enforcing constraints on the asymmetries that can be tolerated.

6.1.3. Saturation

After regeneration only one branch is still active and receive the entire bias current. Until the tail transistor is in saturation, both CM and DM output voltages continue to rise linearly; when also M_0 enters triode, an exponential transient (the resistive tail charging the capacitive load) ends the evolution.

At the very end of the transient, one voltage reaches the supply while the other is still close to $V_{\mathrm{in}}^{\mathrm{cm}}$. If the latch that follows is turned on after this transient has completed, then only one of its input transistors will be on, causing a huge unbalance and making sure the decision of the regenerative circuit goes in the expected direction.

6.1.4. Signal waveforms

Fig. 6.5 shows the most significant waveforms describing the operations of the preamplifier. The bottom plot represents the differential input voltage applied to the circuit. Its values are $\pm$ 5 mV in the two halves of the simulation.

The preamplifier is activated just after instants 0 ns and 250 ns and it is reset after 200 ns and 450 ns. Looking at the v_{o_1} and v_{o_2} waveforms, it is clear how their difference increases over time until a point where the two curves diverge (regeneration, around 90 ns and 340 ns). As already remarked before, on the branch of the highest output voltage, v_x is lower.

The common mode output voltage increases steadily towards the supply, with a slope that decreases slightly at the end of the transient. A reasonable approximation might be to consider it constant, therefore having a lower bound on the duration of the transient. The differential mode output voltage is also varying monotonically.

6.1.5. Mismatch analysis

Mismatches are small deviations of a device parameter from its nominal value. Accounting for such deviations, parameters of topologically symmetric devices can be decomposed into an average and a differential term. The preceding analysis can be considered as based on the average values of the parameters. The differential term is responsible for the coupling of CM and DM modes [19]. This phenomenon can be understood by examining the case of mismatched capacitive loads.

Consider the capacitances expressed as an average and a differential term:

$$C_{L_1} = C_L + \frac{\delta C}{2}$$
$$C_{L_2} = C_L - \frac{\delta C}{2}$$

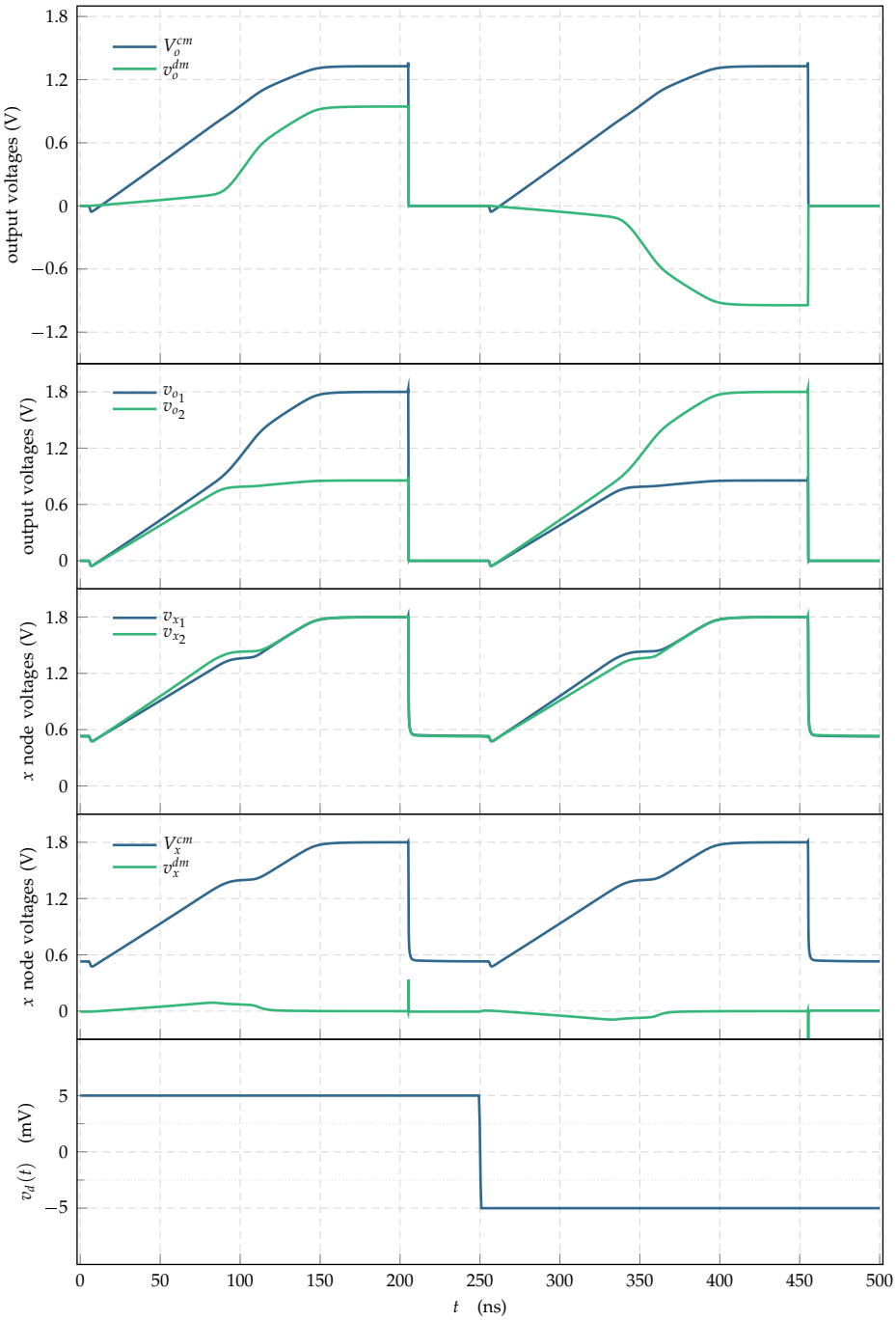

Figure 6.5.: Preamplifier simulated waveforms. Two transients with opposite input differential voltage.

If a constant current I is equally applied to them, the voltage across each capacitor becomes:

$$\begin{aligned} v_{C_1} &= \frac{I}{C_L + \frac{\delta C}{2}} t \\ &\simeq \frac{I}{C_L} t - \frac{I}{C_L} \frac{\delta C}{2C_L} t \\ v_{C_2} &= \frac{I}{C_L - \frac{\delta C}{2}} t \\ &\simeq \frac{I}{C_L} t + \frac{I}{C_L} \frac{\delta C}{2C_L} t \end{aligned}$$

The response due to the average value of the parameter is common to both branches. The mismatch determines a differential component. If such a component is small with respect to the common one, then the operating point of the circuit is not affected. In this example this is meaningless since we are considering passive components and an ideal current source. In an active circuit, where the device parameters depend on the physical quantities in the circuit, this allows us to approximate the response for small differential terms and use superposition to work separately with the common mode and differential mode circuits. The alternative would be to keep track of the coupling by solving the exact differential equations. If the approximation holds, the effort can be spared.

The differential response that stems from a mismatched parameter with a common mode input signal can be moved to the differential mode half circuit. Since the effect depends only on common mode parameters it becomes an independent source.

Conversely, by considering the effects when a differential input is applied to the circuit, the response becomes:

$$\begin{aligned} v_{C_1} &= \frac{i}{C_L + \frac{\delta C}{2}} t \\ &\simeq \frac{I}{C_L} t + \frac{I}{C_L} \frac{\delta C}{2C_L} t \\ v_{C_2} &= \frac{-i}{C_L - \frac{\delta C}{2}} t \\ &\simeq \frac{-I}{C_L} t + \frac{I}{C_L} \frac{\delta C}{2C_L} t \end{aligned}$$

The average parameter value results in a differential component, as desired, while the mismatch generates a common mode response. This becomes an independent source in the common mode equivalent half circuit. However, since this is typically negligible with respect to the original common mode signal, this part of the analysis is not performed.

The technique can be applied to the preamplifier, considering a small variation associated to each parameter in the CM circuit and evaluating the resulting differential response. This, in turn, is placed into the DM circuit as an independent source and

referred to the input in order to determine the equivalent offset voltage in the amplifier transcharacteristic. A few notable cases are shown here, the remaining values in Table 6.1 are derived with the same methodology.

Capacitor C_l

Consider a variation on the load capacitor. It can be represented as a parallel element of value $\frac{\Delta C}{2}$ on one branch, and the opposite on the other. The current it sources is

$$\begin{aligned} i_\Delta &= \frac{\Delta C_l}{2}\frac{\mathrm{d}V_o^{\mathrm{cm}}}{\mathrm{d}t} \\ &= \frac{\Delta C_l}{2}\frac{I_B}{2}\frac{1}{C_l + C_x + 2(C_p + C_{p_x})} \\ &\simeq \frac{\Delta C_l}{2}\frac{I_B}{2}\frac{1}{C_l + C_x} \end{aligned}$$

The effect can be compared to the other DM quantities, by introducing it as an independent generator in the DM equivalent circuit.

If the differential current forced in the branch by v_d is exactly equal to the one just evaluated, then no differential output voltage is generated. In this condition an offset has appeared at the input.

$$\begin{aligned} -g_m\frac{v_d}{2} &= \frac{\Delta C_l}{2}\frac{I_B}{2}\frac{1}{C_l + C_x} \\ v_d^{\mathrm{off}}\Big|_{\Delta C_l} &\simeq \frac{\Delta C_l}{C_l + C_x}\frac{I_B}{2g_m} \end{aligned}$$

Cross-coupled pair V_{off}

Suppose on the gate connection of the cross-coupled pair appears a differential offset voltage V_{off}. Since the XCP is current-biased, the voltage appears as a shift of the source voltage. If this constant differential voltage is so large to change the sign of $v_x^{\mathrm{dm}}(t_1)$, then the wrong decision is taken.

$$\begin{aligned} v_x^{\mathrm{dm}}(t_1) &= \left(\frac{C_l + C_x}{C_l - C_x}\frac{2g_m V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} + 2\sqrt{\frac{(W/L)_1}{(W/L)_2}}\right) v_d \\ &\simeq \left(\frac{2g_m V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} + 2\sqrt{\frac{(W/L)_1}{(W/L)_2}}\right) v_d \\ v_d^{\mathrm{off}}\Big|_{V_{\mathrm{off}}} &\simeq \frac{V_{\mathrm{off}}}{\frac{2g_m V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} + 2\sqrt{\frac{(W/L)_1}{(W/L)_2}}} \end{aligned}$$

Clock transition

The release of the reset condition may not happen simultaneously on both branches. The variation is modeled differentially by a delay t_d applied to the start of the amplification in the CM circuit. Its differential effect is:

$$v_o^{\mathrm{dm}} = \frac{I_B}{2C_L} t_d$$

since it stems from the charging of one load capacitor at half the bias current.

The input-referred offset is evaluated as the v_d that would result in the opposite differential output voltage reached at the end of the linear growth.

$$\begin{aligned} v_d^{\mathrm{off}}\Big|_{t_d} &= -\frac{I_B^2}{g_m V_{\mathrm{in}}^{\mathrm{cm}}} \frac{C_l - C_x}{4C_l(C_l + C_x)} \\ &= -\frac{I_B^2}{g_m V_{\mathrm{in}}^{\mathrm{cm}}} \frac{t_d}{4C_l} \end{aligned}$$

Table 6.1.: Input-referred offset voltage due to parameter variations. Curly braces are used to compact the table when the same multiplicative factor affects a given term

Component	Parameter	Offset (absolute)
Differential couple	V_{th}	ΔV_{th}
	β	$\frac{\Delta\beta}{\beta} \frac{I_B}{2g_m^{\mathrm{diff}}}$
	R_{ds}	$\frac{V_{\mathrm{in}}^{\mathrm{cm}}}{g_m R_{\mathrm{ds}}} \frac{\Delta R}{R_{\mathrm{ds}}}$
Crossed couple	$\{V_{\mathrm{th}}, \beta, V_{\mathrm{off}}\}$	$\frac{\left\{V_{\mathrm{off}}, \Delta V_{\mathrm{th}}, \frac{\Delta\beta}{\beta}\sqrt{\frac{I_B}{\beta}}\right\}}{\frac{2g_m V_{\mathrm{in}}^{\mathrm{cm}}}{I_B} + 2\sqrt{\frac{(W/L)_1}{(W/L)_2}}}$
Capacitances	C_l, C_x	$\frac{\{\Delta C_l, \Delta C_x\}}{C_l + C_x} \frac{I_B}{2g_m^{\mathrm{diff}}}$
	t_{soc}	$\frac{I_B^2}{g_m V_{\mathrm{in}}^{\mathrm{cm}}} \frac{t_d}{4C_l}$

6.2. Regenerative latch

The output of the preamplifier is coupled to the regenerative latch through a couple of parallel transistors that unbalance the internal nodes of the latch by injecting unequal currents.

Modeling the inverters as constant transconductances G_m driven by the voltage on the opposite branch and capacitively loaded, the behaviour of the output voltages can be derived. Performing the analysis with respect to the common and differential modes, the descriptive equations are:

$$V_o^{\mathrm{cm}} = V_o^{\mathrm{cm}}(0)\exp\left(-\frac{G_m}{C_l}t\right) + \frac{I_{cm}}{G_m}\left[1-\exp\left(-\frac{G_m}{C_l}t\right)\right]$$

$$v_o^{\mathrm{dm}} = v_o^{\mathrm{dm}}(0)\exp\left(\frac{G_m}{C_l+2C_p}t\right) + \frac{i_{dm}}{G_m}\left[1-\exp\left(\frac{G_m}{C_l+2C_p}t\right)\right]$$

The result shows a CM voltage that decays exponentially, while the DM signal grows with the same time constant, up to the point where the devices change operating region. The differential output response, ideally, should only stem from the injected currents. However, the fact of having an exponential growth implies careful evaluation of possible noise injected in the output nodes.

An interesting perspective on the evolution of the CM and DM output voltages of a simpler regenerative latch is provided [20]. The model for the latch analyzed here has been compared to the more simple one, however, the results have not been included. What has been observed is that the curves get closer to the middle of the plot, as the injected current speeds up the transient of both modes.

6.2.1. Mismatch considerations

The latch is cascaded to the dynamic preamplifier analyzed previously. If the preamplifier is given enough time to reach saturation of one output, then the matching properties of the latch are not critical, since the output differential mode voltage is around one threshold voltage.

If for some reason, the transient is halted in advance, for example to reduce the power consumption by reducing the on time of the structure, then the input referred offset induced by mismatches in the structure of the latch might become important. The same reasoning already described for the preamplifier, namely the evaluation of the coupling of the CM and DM modes, can be applied here.

One additional mismatch term that is unique to the latch is the one referred to the trip points of the embedded inverters.

7. Mismatch models

Anytime a circuit has to be realized, its elementary components will be affected by errors due to variations of the process parameters. These errors can be of two natures: stochastic, if they can be characterized statistically as instances of some random variable with a given probability density function, or deterministic if they manifest themselves as a gradient along the wafer.

In general, a circuit should be designed to withstand such nonidealities, but in case of extremely accurate operations, the consequences of mismatched parameters might be significant. Therefore the need to quantify such effects at design time, modeling the properties of the technology in use and adjusting the component sizes so that those effects are under control.

The characterization of the devices a given technology can offer is based on several measurements conducted on test structures where the devices have different areas, orientation and distance [21]. However a model interpolating the measured points is required. In the scientific literature several models have been proposed to describe the mismatch of device parameters on a silicon die. They can be classified broadly as physics-based and statistical models. While the former preserve a link to real parameters under the control of the designer, the latter might result in better accuracy, though at the cost of a difficult mapping of the model parameters with real world properties.

Among the physics-based models, the one proposed by Pelgrom et al. in [22] has gradually become the most popular in the design community, in particular for the simplicity of its description. However, the assumptions on which it is based do not allow the treatment of layout styles that decompose the devices in smaller, intertwined sub-elements with the only effects of interdigitation and centroid structures observable on the deterministic component, i.e. long distance gradients.

Since the stochastic component of mismatch could have a spatial correlation larger than the dimensions of one device, ad-hoc positioning of the components might determine better-than-expected performance. Pelgrom's model fails to predict this, providing a pessimistic estimate of the resulting behaviour.

A statistical model that is able to manage the short-range correlation of the stochastic parameter variations comes from Conti [23]. The description assumes a gaussian autocorrelation function for the mismatched parameter and was initially derived to solve the problem observed with Pelgrom's model, of having a variance that grew indefinitely with the distance among identical elements.

This section will try to derive the expected mismatch when interdigitated structures are considered. At first, a brief derivation of Pelgrom's model will be described, showing

where it fails to account for the layout style. Then Conti's model will be introduced in its final form, as presented in the original paper. Its approximated expression for small distances among the elements will be derived, so that the expected mismatch in interdigitated structures can be described analytically, in a easy-to-use formula. Finally, several layout techniques will be compared, in order to define which is the most suitable to the designer's needs.

7.1. Pelgrom's physics-based model

The derivation follows closely the one in [21], with additional comments included for clarity. The model considers a generic parameter P having a spatial dependency on coordinates (x,y). A device occupying some area will be characterized by the average value of P(x,y) over that area. Two devices with identical geometries can therefore be described by different average values:

$$\Delta P_{12} = \frac{1}{S}\left[\int_{S_2} P(x,y)dS - \int_{S_1} P(x,y)dS\right] = G(x,y) * P(x,y),$$

where the last term interprets the difference as a convolution of the mismatch-generating process $P(x,y)$ with a double rectangular function $G(x,y)$:

$$G(x,y) = \frac{1}{A} \cdot \begin{cases} -1 & x,y \in \text{device 1} \\ 1 & x,y \in \text{device 2} \\ 0 & \text{elsewhere} \end{cases}$$

Convolution becomes a product in Fourier domain:

$$\Delta\mathscr{P}(\omega_x,\omega_y) = \mathscr{G}(\omega_x,\omega_y)\mathscr{P}(\omega_x,\omega_y).$$

Moreover, the variance of the stochastic parameter ΔP is equivalent to its power therefore

$$\sigma^2_{\Delta P} = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\mathscr{G}(\omega_x,\omega_y)\right|^2\left|\mathscr{P}(\omega_x,\omega_y)\right|^2 d\omega_x d\omega_y$$

Given a geometry, typically consisting of rectangular elements arranged on the plane, and a description of the mismatch source in terms of spatial frequencies, the variance of the parameter differences characterizing identical devices can be computed.

Up to now, the model is general enough to manage any geometry and mismatch description. The assumption used by the authors to derive its formula is to consider $P(x,y)$ as spatial white noise, with correlation distance of the values at any given point much smaller than the transistor dimensions. Under the previous assumption, the

lowest frequency of the mismatch generating process is much larger than $1/W$ and $1/L$, therefore the formula can be approximated as

$$\sigma^2_{\Delta P} = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\mathscr{G}(\omega_x,\omega_y)\right|^2\left|\mathscr{P}(0,0)\right|^2 d\omega_x d\omega_y \tag{7.1}$$
$$\approx \frac{\left|\mathscr{P}(0,0)\right|^2}{2A} \stackrel{\text{def}}{=} \frac{A_P^2}{WL}.$$

Mismatch is a function of the device area, independently on how the area is built. Two full rectangles placed in proximity of one another or each of them being subdivided and the elementary devices arranged like a chess board, if resulting in the same total area, would provide the same expected mismatch.

Eventually considering a distance-dependent factor in $P(x,y)$, the term linking mismatch to the relative positioning of the devices can be derived, with the complete description of the mismatch being:

$$\sigma^2_{\Delta P} = \frac{A_P^2}{WL} + S_P^2 D^2.$$

In the original paper, the authors actually report the result for a cross coupled geometry, in which, however, each sub-element is as big as the original device. The reduction in mismatch variance, therefore, stems from the increase in the total area, and not on the layout style.

The assumption that prevents the model from being used to evaluate the effect of layout on the performance of symmetric devices is the one that considers the correlation distance of the mismatch-generating process much smaller than the device dimensions, resulting in a constant term inside the integral in (7.1). Before the introduction of the assumption, the model is valid under more general conditions and therefore, we suggest, by using a different description as, for example, a Gaussian profile, a meaningful expression could be derived.

Further work is required, which has not been carried on having found a different model that seemed to solve the problem, though being more oriented towards numerical evaluations than paper-and-pencil calculations.

7.2. Conti's statistical model

This statistical model derived in [23] is again defined on rectangular devices. A random variable $\hat{P}$ characterizes one of the device properties. It results from the spatial average of a process parameter $P(x,y)$, whose value depends on the spatial coordinates:

$$\hat{P} = \frac{1}{A}\int_S P(x,y)dS$$

The model is based around a Gaussian autocorrelation function for the stochastic process P:

$$R_P\left(\tau_x, \tau_y\right) = a_P \exp\left[-K_P^2\left(\tau_x^2 + \tau_y^2\right)\right]$$

where a_P defines the maximum value of the function and K_{PP} the decay constant, also referred to as correlation distance. The autocorrelation defined this way is based on the Euclidean distance between points. It peaks when distance is null, is close for nearby elements small distances and vanishes as they are separated.

The model derived in the original paper for the variance of the mismatch in parameter P is then:

$$\sigma^2_{\Delta\hat{P}} = 2\left[\Omega\left(L, W, 0, 0, a_P, K_P\right) - \Omega\left(L, W, D_x, D_y, a_P, K_P\right)\right]$$

where

$$\Omega\left(L, W, d_x, d_y, a_P, K_P\right) = \frac{a_P}{(2K_{PP}^2 LW)^2} A(L, d_x, K_P) A(W, d_y, K_P)$$

and

$$\begin{aligned} A(r, d, K_P) =& \sqrt{\pi} K_P \left[(d-r)\text{erf}\left[K_P(d-r)\right] - 2d\text{erf}\left[K_P d\right] + (d+r)\text{erf}\left[K_P(d+r)\right]\right] \\ &+ \exp\left[-K_P^2(d-r)^2\right] - 2\exp\left[-K_P^2 d^2\right] + \exp\left[-K_P^2(d+r)^2\right] \end{aligned}$$

If the devices are partitioned into n_d smaller elements, the variance becomes

$$\sigma^2_{\Delta\hat{P}} = \frac{2}{n_d^2}\sum_{l=1}^{n_d}\sum_{m=1}^{n_d}\left[\Omega\left(L, W/n_d, D_{x_{lm}}^{(1)(1)}, D_{y_{lm}}^{(1)(1)}, a_P, K_P\right) - \Omega\left(L, W/n_d, D_{x_{lm}}^{(1)(2)}, D_{x_{lm}}^{(1)(2)}, a_P, K_P\right)\right]$$

This description is very powerful, since it can be applied to any arrangement of the elements, taking into account the mutual distance between them. Depending on the layout style, the mutual distances take different forms, therefore this expression can be easily applied to evaluate how particular layout strategies affect the behaviour of nominally-identical devices in symmetric structures.

The comparison between the physics-based and statistical models is depicted in Fig. 7.1. The most striking difference is the value of mismatch for low distances, which is much lower in Conti's model. This result comes from the increased correlation of stochastic device parameters when they are placed in close proximity.

Fig.7.2 shows how the mismatch for two adjacent devices vary with W and L. Since the curves are non monotonic, we suppose that a different domain might be more

suitable. Indeed, using as coordinates the area and aspect ratio of the devices, the surface becomes convex (Fig.7.3).

The constant mismatch contours of the surface in Fig.7.3 have been plotted in Fig.7.4, together with the constant length lines. Choosing the desired channel length, W is defined by the allowed mismatch.

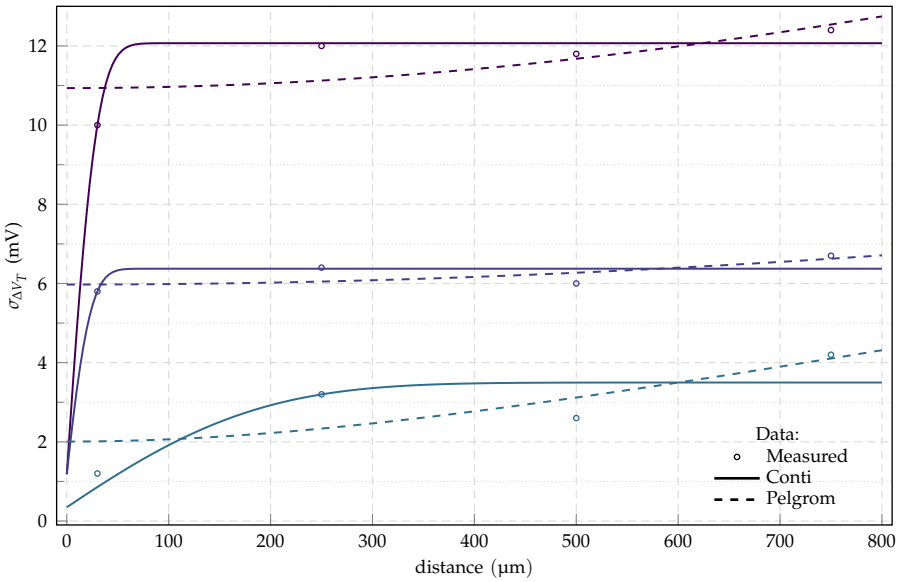

Figure 7.1.: Comparison of Pelgrom's and Conti's interpolation on the mismatch data used by Pelgrom in his original paper.

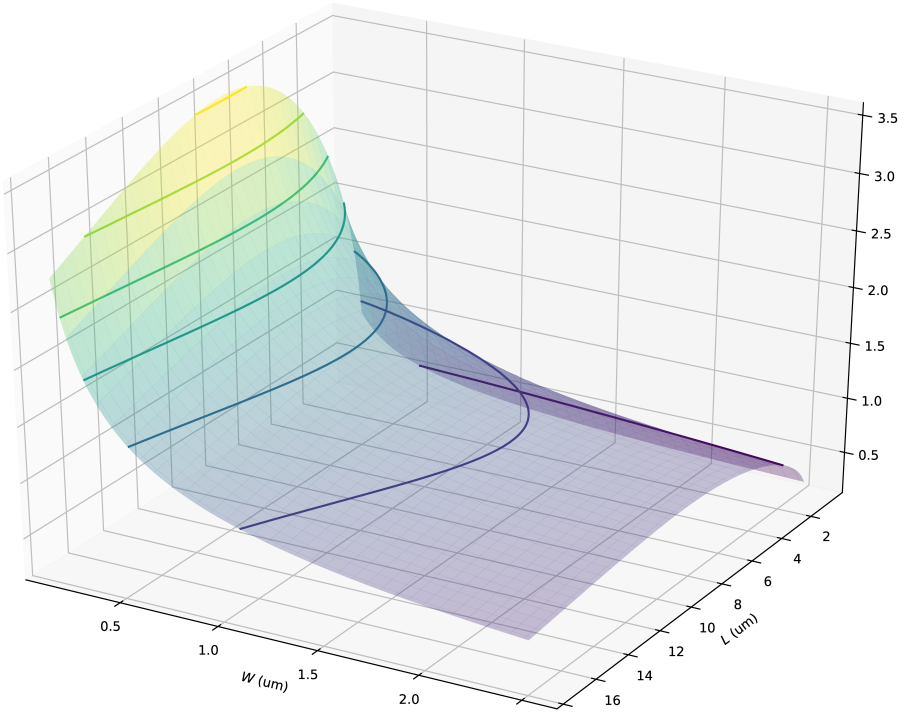

Figure 7.2.: Mismatch $\sigma_{\Delta V_{th}}$ (mV) in terms of W and L of the devices.

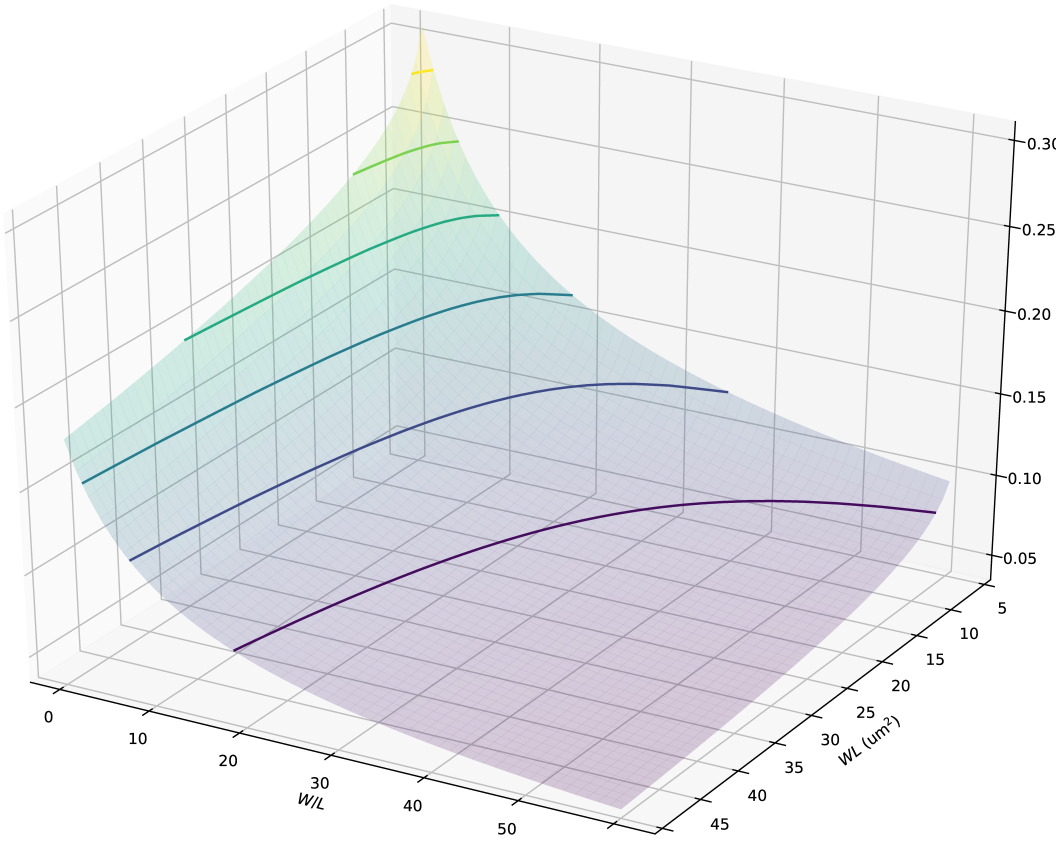

Figure 7.3.: Mismatch $\sigma_{\Delta V_{th}}$ (mV) in terms of area and aspect ratio of the devices.

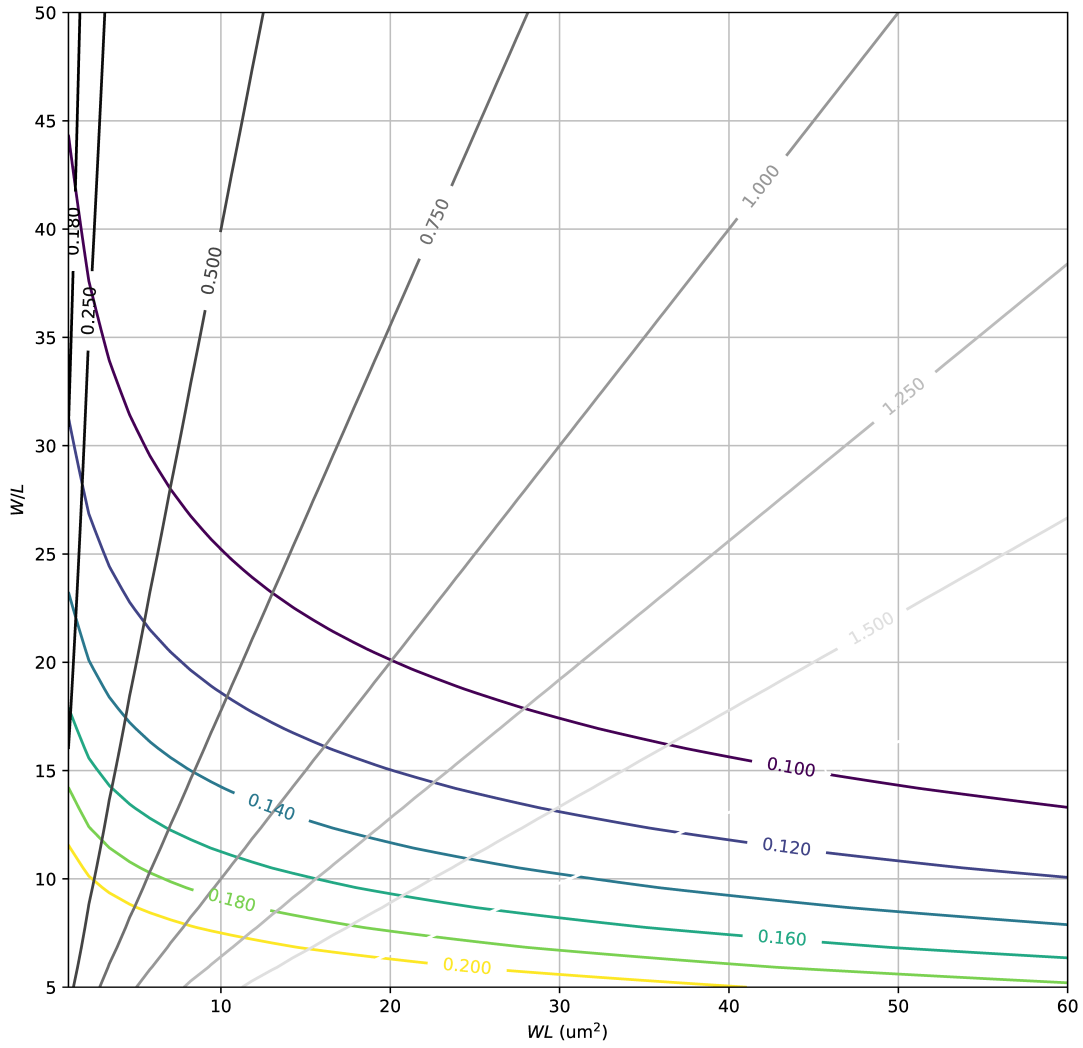

Figure 7.4.: (Coloured) Contour plot of $\sigma_{\Delta V_{th}}$ (mV) in terms of area and aspect ratio of the devices. (Greyscale) Constant L (µm) curves

7.2.1. Approximated model for short distance mismatch

In order to obtain some simple relationship on the layout dependence of mismatch, the functions involved in the definition of the model have been approximated around 0, starting from the innermost functions.

The $\mathrm{erf}(x)$ and $\exp(x)$ functions have to be approximated to fourth order to get to a meaningful result:

$$\mathrm{erf}(x) \sim \frac{2x}{\sqrt{\pi}} - \frac{2x^3}{3\sqrt{\pi}} + \mathrm{o}\left(x^4\right)$$
$$\exp(x) \sim 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \mathrm{o}\left(x^4\right)$$

The model expressions thus become:

$$A(r,d,k) \sim 2k^2r^2\left(1 - k^2\left(d^2 + \frac{r^2}{6}\right)\right)$$
$$\Omega = a\left(1 - k^2\left(d_x^2 + \frac{1}{6}L^2\right)\right)\left(1 - k^2\left(d_y^2 + \frac{1}{6}W^2\right)\right) \tag{7.2}$$

7.3. Effects of layout style on mismatch

Assuming a linear arrangement of the elements, only one of the distance terms in Ω will be different from zero at a given time. Therefore the function can be rewritten as:

$$\Omega = a\left(1 - k^2\left(d_\xi^2 + \frac{1}{6}\lambda^2\right)\right)\left(1 - k^2\left(\frac{1}{6}\omega^2\right)\right),$$

where d_ξ is the distance along the alignment axis, λ is the length of the elementary rectangle along said axis and ω is the length in the orthogonal direction. If the elements are aligned along x, with horizontal dimension L and vertical one W, then the mapping is:

$$d_\xi = d_x$$
$$\omega = W$$
$$\lambda = L$$

The variance requires the computation of the difference

$$\Omega^{(1,1)} - \Omega^{(1,2)} = a\left(1 - k^2\left(\frac{1}{6}\omega^2\right)\right)\left(-k^2\left({d_\xi^{(1,1)}}^2 - {d_\xi^{(1,2)}}^2\right)\right)$$

Consider that the distances between the elements of index l and m can be expressed, depending on the device to which they belong, as:

$$d_{\xi}^{(1,1)} = 2|l-m|d_u$$
$$d_{\xi}^{(1,2)} = 2|l-m+1|d_u$$

where $d_u = d_s + \lambda$ is the unitary distance between two similar corners of the device geometry and d_s is the actual spacing (empty space) between the elements Therefore the previous difference can be further developed, becoming

$$\Omega^{(1,1)} - \Omega^{(1,2)} = a\left(1 - k^2\left(\frac{1}{6}\omega^2\right)\right)\left(-4k^2\left(2(l-m)+1\right)\right)$$

Finally:

$$\sigma^2 = 8ak^2d_u^2\left(1 - \frac{1}{6}k^2\omega^2\right)$$

In the case of x-aligned elements, without spacing among them, this becomes

$$\sigma^2 = 8ak^2L^2\left(1 - \frac{1}{6}k^2\frac{W^2}{n_d^2}\right)$$

And it varies as n_d^{-2}.

The same sequence of operations can be carried out in the case of the sub-elements positioned in two columns, alternating the device order in each one. This layout corresponds to a common centroid when n_d is even. Both distances have to be considered. They can be expressed as:

$$d_{\xi}^{(1,1)} = d_{\xi,u} \cdot \begin{cases} 1 & \text{if } l,m \text{ both even or odd} \\ 0 & \text{else} \end{cases}$$
$$d_{\xi}^{(1,2)} = d_{\xi,u} \cdot \begin{cases} 0 & \text{if } l,m \text{ both even or odd} \\ 1 & \text{else} \end{cases}$$
$$d_{\omega}^{(1,1)} = d_{\omega}^{(1,2)} = d_{\omega,u}|l-m|$$

As a result, the distances

$$d_{\xi}^{(1,1)} - d_{\xi}^{(1,2)} = d_{\xi,u}(-1)^{l+m}$$
$$d_{\omega}^{(1,1)} - d_{\omega}^{(1,2)} = 0$$

Computing $\Omega^{(1,1)} - \Omega^{(1,2)}$ starting from the complete expression for Ω (7.2) and substituting the distances with the terms just derived, we find:

$$\sigma^2 = \begin{cases} a\left(\frac{k^2LW}{n_d^2}\right)^2 & \text{if } n_d \text{ even} \\ 2a\left(\frac{kL}{n_d}\right)^2\left(1 - \frac{(kW)^2}{2} + \frac{1}{3}\left(\frac{kW}{n_d}\right)^2\right) & \text{if } n_d \text{ odd} \end{cases} \tag{7.3}$$

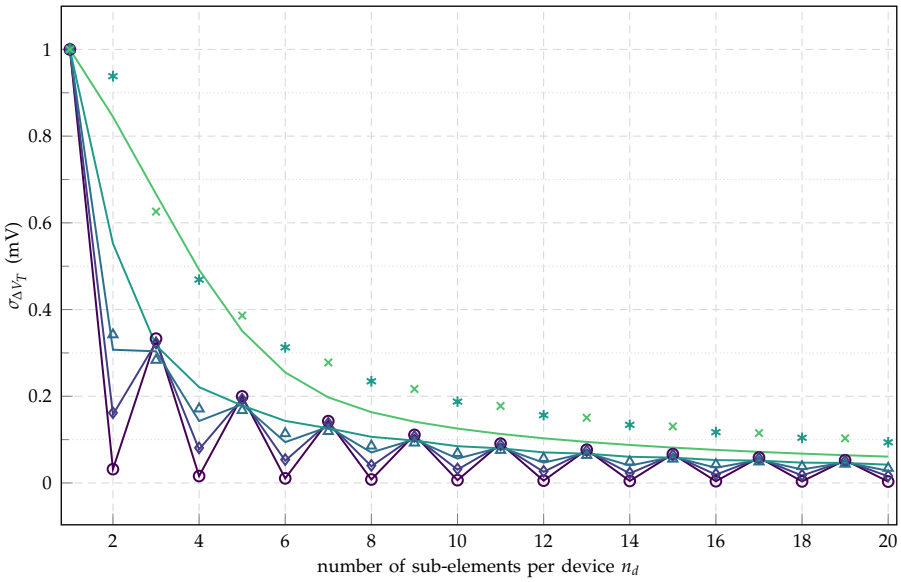
1
0.8
0.6
0.4
0.2
0
$\sigma_{\Delta V_T}$ (mV)
2
4
6
8
10
12
14
16
18
20
number of sub-elements per device n_d

Figure 7.5.:

8. Conclusion

This thesis has been focused on the design of a low power Analog-to-Information architecture. After an initial review of the underlying theory of Compressed Sensing, the modifications applied to a traditional charge-redistribution successive-approximation-register A/D converter have been discussed.

The identification of the limitations determined by technological factors has lead to the introduction of additional circuital elements (the leakage compensator in particular) as well as more abstract modification (the design of the sensing matrix) so that the performance of the real system could closely match an ideal CS-based signal processing chain.

Several non-idealities have been modeled, with the intention of guiding the designer in more appropriate choices both on the circuit topologies to be employed and proper device sizing. The design constraints derived in the thesis, applied to the architecture under consideration, have been validated through algorithmic and circuital simulation, highlighting the effective performance gains.

Future work would involve further modeling of some of the phenomena whose analysis has been left incomplete. Then all circuit elements have to be sized, again taking great consideration of all possible non-idealities and culminating in the layout of the entire structure. Finally measurements should be performed on the actual chip, validating hopefully the modeling-driven design choices. Taking advantage of the extensive analysis effort started with this thesis, the design should be straightforward, leading to an effective use of the technology and resulting in an efficient acquisition system.

Bibliography

[1] M. Unser, "Sampling-50 years after Shannon," vol. 88, no. 4, pp. 569–587.

[2] S. Greco and P. Valabrega, *Algebra lineare*. Libreria editrice universitaria Levrotto & Bella, oCLC: 800078771.

[3] S. Theodoridis, *Machine Learning*. Elsevier.

[4] "Compressed Sensing: Theory and Applications."

[5] M. Mangia, F. Pareschi, V. Cambareri, R. Rovatti, and G. Setti, *Adapted Compressed Sensing for Effective Hardware Implementations*. Springer International Publishing.

[6] A. Mishra, F. N. Thakkar, C. Modi, and R. Kher, "Selecting the Most Favorable Wavelet for Compressing ECG Signals Using Compressive Sensing Approach," in *2012 International Conference on Communication Systems and Network Technologies*. IEEE, pp. 128–132.

[7] S. G. Mallat, S. G. Mallat, and Elsevier Science (Firm), *A Wavelet Tour of Signal Processing: The Sparse Way*. Connexions, Rice University, oCLC: 707720492.

[8] D. Bortolotti, M. Mangia, A. Bartolini, R. Rovatti, G. Setti, and L. Benini, "Energy-Aware Bio-Signal Compressed Sensing Reconstruction on the WBSN-Gateway," vol. 6, no. 3, pp. 370–381.

[9] P. E. McSharry, G. D. Clifford, L. Tarassenko, and L. A. Smith, "A dynamical model for generating synthetic electrocardiogram signals," vol. 50, no. 3, pp. 289–294.

[10] M. Pelgrom, *Analog-to-Digital Conversion*. Springer International Publishing.

[11] F. Pareschi, P. Albertini, G. Frattini, M. Mangia, R. Rovatti, and G. Setti, "Hardware-Algorithms Co-Design and Implementation of an Analog-to-Information Converter for Biosignals Based on Compressed Sensing," vol. 10, no. 1, pp. 149–162.

[12] K. Bernstein, A. R. Bonaccio, J. A. Fifield, A. P. Haar, S. C. Ho, T. B. Hook, M. A. Soma, and S. D. Wyatt, "Leakage compensation circuit," patentus 6 956 417B2.

[13] K.-K. Kim and Y.-B. Kim, "A 32nm and 0.9V CMOS Phase-Locked Loop with Leakage Current and Power Supply Noise Compensation," vol. 7, no. 1, pp. 11–19.

[14] L. S. Y. Wong, S. Hossain, and A. Walker, "Leakage current cancellation technique for low power switched-capacitor circuits," in *Proceedings of the 2001 International Symposium on Low Power Electronics and Design*, pp. 310–315.

[15] G. Marro, *Controlli automatici*. Zanichelli, oCLC: 860501224.

[16] V. Vorperian, *Fast Analytical Techniques for Electrical and Electronic Circuits*. Cambridge University Press.

[17] R. Middlebrook, "Low-entropy expressions: The key to design-oriented analysis," in *Proceedings Frontiers in Education Twenty-First Annual Conference. Engineering Education in a New World Order*. IEEE, pp. 399–403.

[18] Y. Zhang, E. Bonizzoni, and F. Maloberti, "A 10-b 200-kS/s 250-nA Self-Clocked Coarse–Fine SAR ADC," vol. 63, no. 10, pp. 924–928.

[19] R. D. Middlebrook, *Differential Amplifiers*. John Wiley & Sons Inc.

[20] A. Abidi and H. Xu, "Understanding the regenerative comparator circuit," in *Proceedings of the IEEE 2014 Custom Integrated Circuits Conference*. IEEE, pp. 1–8.

[21] J. A. Croon, H. E. Maes, and W. Sansen, *Matching Properties of Deep Sub-Micron MOS Transistors*, oCLC: 1012540433.

[22] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of mos transistors," vol. 24, no. 5, pp. 1433–1439, oCLC: 5872080160.

[23] M. Conti, P. Crippa, S. Orcioni, and C. Turchetti, "Layout-based statistical modeling for the prediction of the matching properties of MOS transistors," vol. 49, no. 5, pp. 680–685.