EXPLOITING FLOATING-GATE TRANSISTOR PROPERTIES IN ANALOG AND MIXED-SIGNAL CIRCUIT DESIGN

A Dissertation
Presented to
The Academic Faculty

By

Erhan Özalevli

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Electrical and Computer Engineering

School of Electrical and Computer Engineering
Georgia Institute of Technology
December 2006

EXPLOITING FLOATING-GATE TRANSISTOR PROPERTIES IN ANALOG AND MIXED-SIGNAL CIRCUIT DESIGN

Approved by:

Dr. Paul E. Hasler, Advisor
Professor, School of ECE
Georgia Institute of Technology
Atlanta, GA

Dr. Charles M. Higgins
Professor, School of ECE
The University of Arizona
Tucson, AZ

Dr. David V. Anderson
Professor, School of ECE
Georgia Institute of Technology
Atlanta, GA

Dr. Alan Doolittle
Professor, School of ECE
Georgia Institute of Technology
Atlanta, GA

Dr. Farrokh Ayazi
Professor, School of ECE
Georgia Institute of Technology
Atlanta, GA

Date Approved: July 2006

DEDICATION

To my family...

ACKNOWLEDGEMENTS

I would like to thank my family for their endless support and love through all my endeavors.

I wish to express my sincere gratitude to my advisor Dr. Hasler for his support throughout my PhD. I am also grateful to Dr. Higgins, Dr. Anderson, and Dr. Ayazi for serving in my thesis defense committee.

Also, I would like to thank all the members of the CADSP Lab for a pleasant and friendly atmosphere, especially I am very thankful to Guillermo, Shakeel, Venkatesh, Chris, Kofi, Thomas, Gail, David, Ryan, Degs, Huseyin, and Jenny for their friendship and support. Lastly, I would like to thank Serdar, Koray, Yakup, Gunay, Zafer, and Menderes for their friendship and good company...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

With the downscaling trend in CMOS technology, it has been possible to utilize the advantages of high element densities in VLSI circuits and systems. This trend has readily allowed digital circuits to predominate VLSI implementations due to their ease of scaling. However, high element density in integrated circuit technology has also entailed a decrease in the power consumption per functional circuit cell for the use of low-power and reconfigurable systems in portable equipment.

Analog circuits have the advantage over digital circuits in designing low-power and compact VLSI circuits for signal processing systems. Also, analog circuits have been employed to utilize the wide dynamic range of the analog domain to meet the stringent signal-to-noise-and-distortion requirements of some signal processing applications. However, the imperfections and mismatches of CMOS devices can easily deteriorate the performance of analog circuits when they are used to realize precision and highly linear elements in the analog domain. This is mainly due to the lack of tunability of the analog circuits that necessitates the use of special trimming or layout techniques.

These problems can be alleviated by making use of the analog storage and capacitive coupling capabilities of floating-gate transistors. In this research, tunable resistive elements and analog storages are built using floating-gate transistors to be incorporated into signal processing applications. Tunable linearized resistors are designed and implemented in CMOS technology, and are employed in building a highly linear amplifier, a transconductance multiplier, and a binary-weighted resistor digital-to-analog converter. Moreover, a tunable voltage reference is designed by utilizing the analog storage feature of the floating-gate transistor. This voltage reference is used to build low-power, compact, and tunable/reconfigurable voltage-output digital-to-analog converter and distributed arithmetic architecture.

CHAPTER 1

EXISTING APPROACHES IN ANALOG AND MIXED-SIGNAL CIRCUIT DESIGN

Maintaining the signal integrity and precision through the signal processing path is one of the most challenging issues in analog and mixed-signal circuit design. To achieve this, analog and mixed-signal circuits are generally designed to preserve the accuracy and precision in the signal amplitude and time while processing them. On the other hand, digital circuits process the information in two amplitude states of a bit during a predefined time interval, thus the accuracy in the signal amplitude is not the main constraining issue for digital circuits. Therefore, the demands of analog and mixed-signal circuits from the process technology are different from that of digital circuits.

The scaling in the process technology has enabled designers to obtain high element densities with digital circuits. However, this scaling trend has imposed different design challenges for analog and mixed-signal circuits and the cost-effective CMOS integration. Especially as the supply voltage is decreased due to the technology scaling it has become more difficult to process the signals in the analog domain with the reduced voltage headroom. In addition, the relative parametric variations has increased with the scaling in the process technologies [1], making the linearity, noise, and distortion issues become more difficult to overcome in analog and mixed-signal circuits.

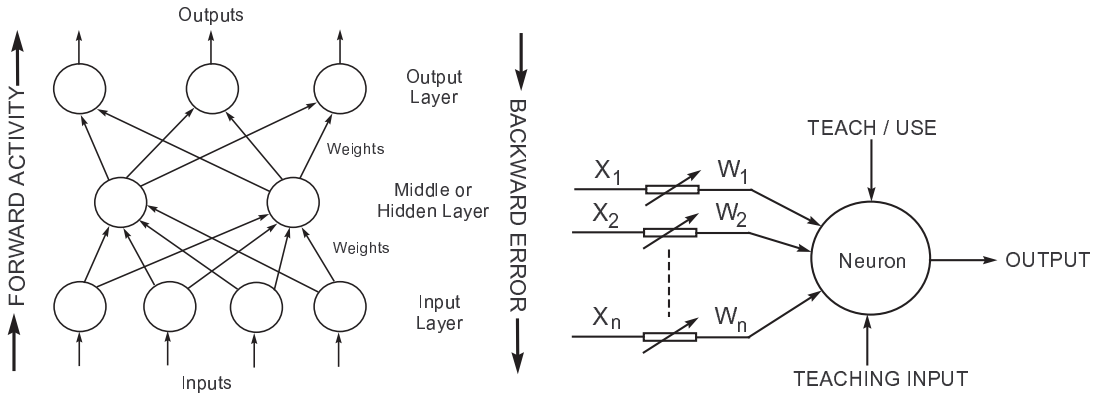
While being less prone to the device imperfections, digital circuits also offer design flexibility and reconfigurability. However, it is necessary for some applications to use special-purpose digital circuitry since reconfigurable digital-signal processing circuits are generally large and power-hungry [2], [3]. The multiplication and addition operations are the repetitive functions frequently used in signal-processing systems. Even for custom digital circuits, their digital implementations cause increase in the total die area and power consumption making it difficult to realize the low-power digital circuit implementations

of the signal-processing systems. In contrast, the area and power consumption associated with the addition and multiplication operations can be easily optimized with analog and mixed-signal circuit implementations. Moreover, a variety of design strategies has been employed for analog and mixed-signal circuits to achieve reconfigurability and tunability and to deal with the device mismatches and imperfections. For instance, tunable resistors are incorporated into artificial neural networks (ANN) to set and tune the synaptic weights. Similarly, linearization techniques are employed for highly linear amplifiers and multipliers to increase the circuit linearity and minimize the signal distortion. In data converters, a variety of calibration methods are utilized to alleviate the device imperfections. Furthermore, switched-capacitor and switch-current techniques are employed for analog and mixed-signal circuits to achieve the reconfigurability and tunability. In the subsequent sections, these techniques and their circuit implementations will be summarized.

1.1 Tunability in artificial neural network (ANN) systems

ANN is an information processing system inspired by the biological nervous systems. It consists of highly interconnected processing elements configured to solve specific problems and to achieve certain tasks. Adaptive ANN systems learn by example, and like it is the case for biological systems they adjust their synaptic weights and connections to adapt to their changing environments.

Figure 1a illustrates a typical artificial neural network architecture [4], where the inputs are usually binary, and the connections between the input layer and the middle or hidden layer contain the weights. These weights are generally determined by training the system. In addition, the middle layer processes the weighted inputs and sums them. The output is created based on the transfer function of the system. This transfer function can be a sigmoid function, which varies from 0 to 1 for a range of inputs. The connections between the middle and output layer also have weights, and the output layer contains the transfer function of the system.



(a) (b)

Figure 1. (a) Typical artificial neural network setup [4]. (b) McCulloch and Pitts neuron model [5]. The inputs are weighted so that the effect that each input has at decision making is dependent on the weight of the particular input

Moreover, a neuron model by McCulloch and Pitts [5] is depicted in Figure 1b. In this model, the inputs are weighted so that their effect at decision making is dependent on the weight of a particular input. These weighted inputs are then added together and if they exceed a pre-set threshold value, the neuron fires. This neuron model has the ability to adapt to a particular situation by changing its weights and/or threshold. This has been achieved by employing algorithms such as the back error propagation and the Delta rule.

The synaptic weights in ANN systems can be implemented in CMOS processes by using resistors [6]. The resistors in such applications can be designed and made tunable by exploiting the CMOS transistor properties. While the linearity is one of the most important metrics used to design tunable CMOS resistors, they are usually built based on the specifications imposed by their application. Therefore, depending upon the application, the CMOS resistors are generally required to be highly linear, area and power efficient, and to have a wide tuning/operating range. The compactness, power efficiency, and tuning range are the primary concerns for ANN systems.

In a standard CMOS technology, linearized tunable CMOS resistors are designed by applying linearization techniques to MOS transistors. These techniques exploit both the

MOSFET's square-law characteristic in the saturation region [7], [8], and its resistive nature in the triode region [9], either separately or in combination [10], [11]. Although the linearization of MOS transistors in the saturation region has been achieved to obtain CMOS resistors with reduced nonlinearity, such as by applying a bias-offset technique [12] or a square-law method [7], these structures generally suffer from channel-length modulation, mobility degradation, and device mismatches. In addition, MOS transistors have been linearized by operating them in the triode region, and using balanced networks [13], [14], [9] or depletion devices [15]. However, the balanced resistor structures are sensitive to the mismatches that cause even-ordered distortion, and to the mobility degradation that results in odd-ordered distortion. Moreover, the tuning range of the linearization technique with depletion devices are strongly limited [16]. Alternative to these approaches, the gate linearization [17] or common-mode strategy [18] can be adopted to a single MOS transistor in the triode region to alleviate the linearity, mismatch, operating-range, and tuning-range issues.

An example of a tunable resistor is a voltage controlled resistor [19] shown in Figure 2a. This resistor is similar to the resistor structures proposed by Rasmussen [20] and Singh [8]. In addition to the mirror transistors, four more MOS transistors and two control voltage sources are used for the design of this resistor. A pair of MOS transistors, M_1 and M_2, is connected as a bilateral resistor while the other pair of MOS transistors, M_3 and M_4, is similarly connected to the middle right of the circuit. These resistors are controlled by the common voltages V_{cp} and V_{cn}.

Furthermore, a tunable CMOS resistor for ANN systems can be also implemented by using the circuit shown in Figure 2b. This circuit is a floating resistor exhibiting positive or negative resistance values depending on its biases [21]. The transistor nonlinearities are cancelled by operating the transistors in their saturation region. The nodes V_X and V_Y are the two terminals of the resistor, and the resistance is inversely proportional to the difference of control voltages V_{C1} and V_{C2}. If V_{C1} is greater than V_{C2}, the circuit operates

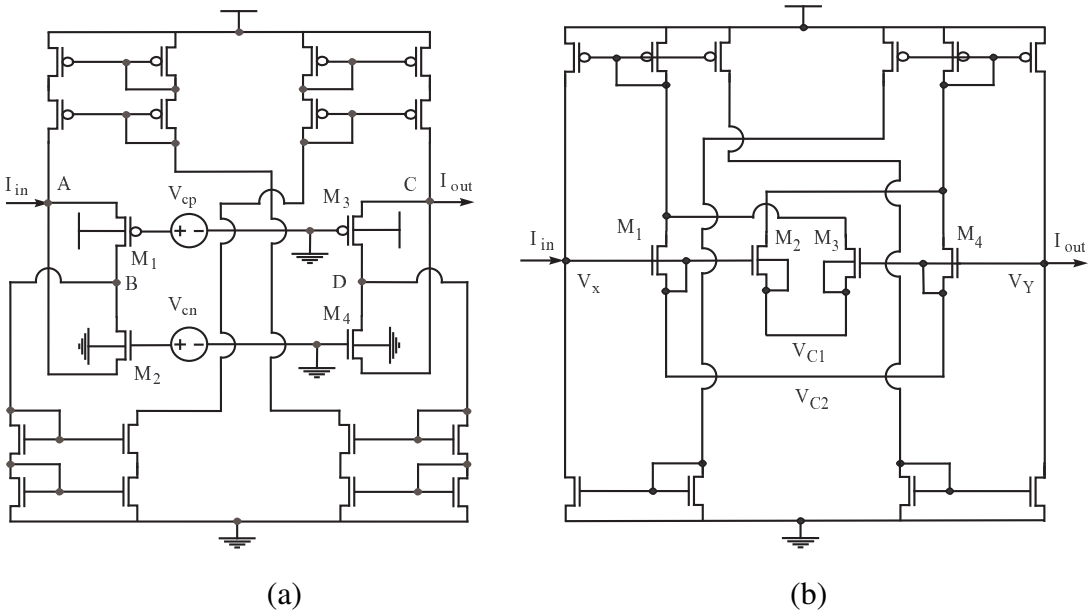


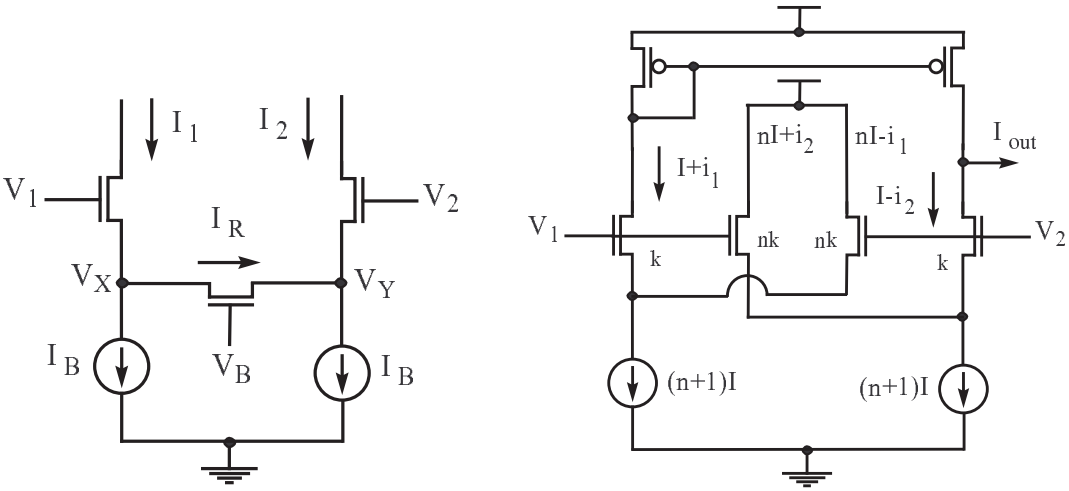
Figure 2. (a) Circuit schematic of the CMOS bilateral linear floating resistor [19]. (b) Circuit diagram of floating resistor [21].

as a resistor circuit with positive resistance. Alternatively, if V_{C2} is greater than V_{C1}, the circuit operates as a resistor circuit with negative resistance.

1.2 Linearity of Highly Linear Amplifier and Multiplier Circuits

Highly linear amplifiers and transconductance multipliers are two of the most versatile analog circuit blocks and are widely used in signal and information processing applications. Highly linear amplifiers are particularly important for the design of data converters and continuous-time filters, and multipliers are essential components of modulators and mixers. The stringent signal-to-noise-and-distortion requirements of these applications usually require highly linear circuits that can handle large signal swings at their inputs/outputs.

The linear range of differential amplifiers can be increased by employing resistive source-degeneration techniques. A single MOS transistor in triode region can be used to serve this purpose [22] as shown in Figure 3a. However, the use of a single transistor alone is not effective due to the fact that MOS transistors in triode region exhibit a large dependence on the common mode of its input signals. Another approach is to use a cross-coupled



(a) (b)

Figure 3. (a)$V - I$ conversion based on a single MOS triode transistor[22]. (b) Circuit realization of the linearized transconductance based on the cross-coupled quad configuration [23].

quad cell that has transistors n times larger than the input transistors and acts as a source follower to create a constant sum of V_{gs} [23] as illustrated in Figure 3b. This topology results in increased power consumption, and the linearity of the amplifier is limited. Similar to the amplifiers, transconductance multipliers implemented with MOS transistors in the triode region suffer not only from mismatch and offset, but also from the MOS transistor nonlinearities which becomes significant for larger input swings [24].

1.3 Design issues of digital-to-analog converters and multi-bit quantizers

Traditional DAC designs are driven by their applications and are generally subject to constraints imposed by the trade-offs between power, speed, resolution, and area. This is especially the case for embedded on chip systems where die area tends to be a major concern. Depending upon the application, accuracy and/or resolution is often sacrificed for reduced area.

1.3.1 Binary-weighted capacitor DAC

Within the Nyquist rate DACs, the binary-weighted capacitor DAC (BWCDAC) allows for obtaining a good accuracy [25]. This DAC architecture, shown in Figure 4, was first presented by McCreary and Gray [26], and implemented by utilizing the scaled capacitors. Although it yields a good accuracy, its binary-weighted capacitor array causes a large element spread and an exponential growth in the total area as the number of bits increases. Also, the achievable resolution and accuracy of this DAC is limited for higher resolutions, since the matching accuracy of the capacitors degrades as the capacitor ratio increases. In order to ease the area and resolution trade-off, DAC architectures based on two stage capacitor arrays [27], and $C-2C$ ladders [28] were proposed. $C-2C$ ladder structure is one of the best area optimization technique for the BWCDAC, since in this case, the area increases linearly with the number of bits and the element spread is only 2. However, the accuracy of this DAC is sensitive to the parasitic capacitances at the capacitor ladder interconnections.

While it is possible to reduce the total area of the BWCDAC by employing different design strategies, the accuracy and the area of this converter is mostly dictated by the capacitor matching. Therefore, it becomes crucial for this kind of converters to have minimized capacitor mismatches. The mismatch between capacitors is caused by the systematic and random errors [29–31]. The area and perimeter of capacitors, capacitor-to-capacitor gap, corner-cutting, and capacitor ratio determine the maximum achievable capacitor matching [32]. The capacitor matching can be improved by employing unit capacitors that have the same perimeter-to-area ratio. Although, a precise capacitor matching (around 0.01%) in modern CMOS processes can be obtained by employing different layout techniques [33], the total capacitor area dictated by the capacitor matching and unit capacitor size increases with these techniques. It has been shown that capacitor mismatch errors can be filtered out by employing sinal processing techniques such as dynamic element matching [34], data-weighted averaging [35], and noise-shaping [36, 37] techniques. These design strategies use digital signal processing techniques to minimize the effect of the mismatch errors in

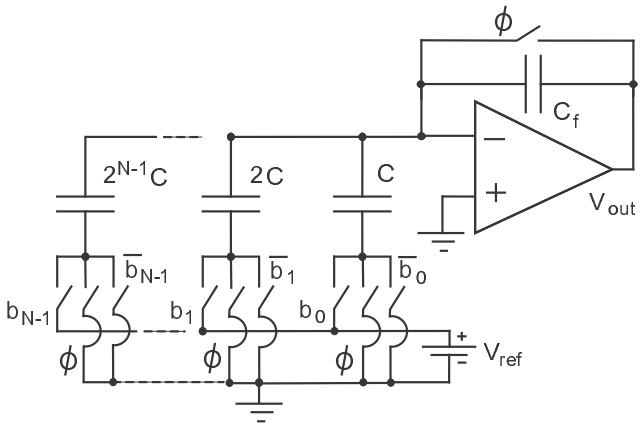


Figure 4. Traditional design of binary-weighted capacitor charge amplifier DAC circuit. C_f is the feedback capacitor and equal to $2^N C$. ϕ is the digital reset signal used to clear the inverting-node of the amplifier.

the frequency range of interest. For that purpose, the sampling rate has to be increased enough to allow for the over-sampling of the input signal.

1.3.2 Multi-bit quantizers using binary-weighted resistor DAC

In multi-bit-per-stage pipelined and sub-ranging converters as well as in oversampling converters, multi-bit quantizers can be successfully employed to improve the overall performance. In pipelined ADCs, the use of multi-bit quantizers decreases the number of stages and reduces the conversion latency. Also, interstage analog signal processing performance can be optimized depending on the accuracy of the sub-stages. Proper selection of stage resolution and use of multi-bit quantizers allow for the optimization of silicon area, power consumption, and conversion speed for resolutions higher than 10 bits [38].

Similarly, multi-bit quantizers are important in building oversampling converters. When designing a converter with a high dynamic range for the low-voltage and low-power applications, the signal swing at the integrator output needs to be lowered, and this requirement can be readily met by employing multi-bit quantizers. Also, increasing the number of bits of the internal quantizer in $\Delta\Sigma$ modulators enables for the reduction of the quantization noise by $6dB$ for each additional bit, and improves the stability of the higher order $\Delta\Sigma$ modulators [39] [40].

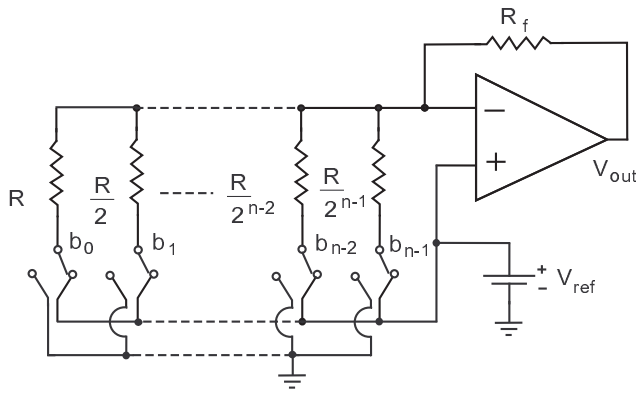


Figure 5. Traditional design of n-bit binary-weighted resistor DAC circuit. R_f is the feedback resistor and b_i is the digital input bit for $i = 0, .., N-1$.

A quantizer can be easily built by using a binary-weighted resistor DAC structure shown in Figure 5. Although this kind of DAC structure can be fast and insensitive to parasitics, it is susceptible to resistor mismatches, which can substantially alter the linearity performance of the converter. Passive resistors in CMOS technologies are typically implemented by utilizing polysilicon, diffusion or well strips. These resistors exhibit around ±0.1% matching accuracy and ±30% tolerance due to device-to-device and lot-to-lot variations in semiconductor fabrication processes [33]. Thin film resistors typically have much better matching accuracy and temperature coefficients, but they are not available in the main stream CMOS processes. The device mismatches and component variations in CMOS processes are generally minimized by employing calibration methods. These calibration techniques include trimming and the use of programmable binary-weighted array. Component trimming is achieved during the test phase of the production by using laser technology. The programming method is used to choose the desired array of elements by blowing fuses. These methods are irreversible and introduce problems over time due to aging, stress, and temperature.

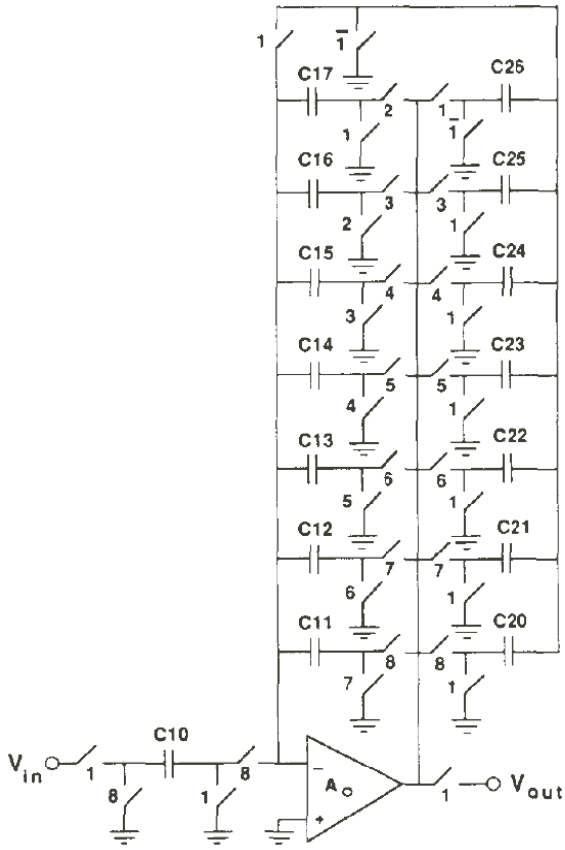


Figure 6. Example of switched-capacitor FIR filters. A general purpose 6th-order direct-form FIR filter by using switched-capacitor technique [43].

1.4 Tunability and reconfigurability in the implementations of the finite impulse response filters

To obtain a programmability in the analog domain, a variety of design strategies has been suggested [41, 42]. The analog and mixed-signal implementations of FIR filters have been generally designed for pre and post-processing applications by employing switched-capacitor and switch-current techniques.

Switched-capacitor techniques are suitable for FIR filter implementations and offer precise control over the filter coefficients. A general purpose of FIR filter implementation based on the switched-capacitor technique is illustrated in Figure 7a. These techniques pose different design challenges depending upon the implementation. To avoid the power and speed trade-off in the switched-capacitor FIR filter implementations, a transposed FIR

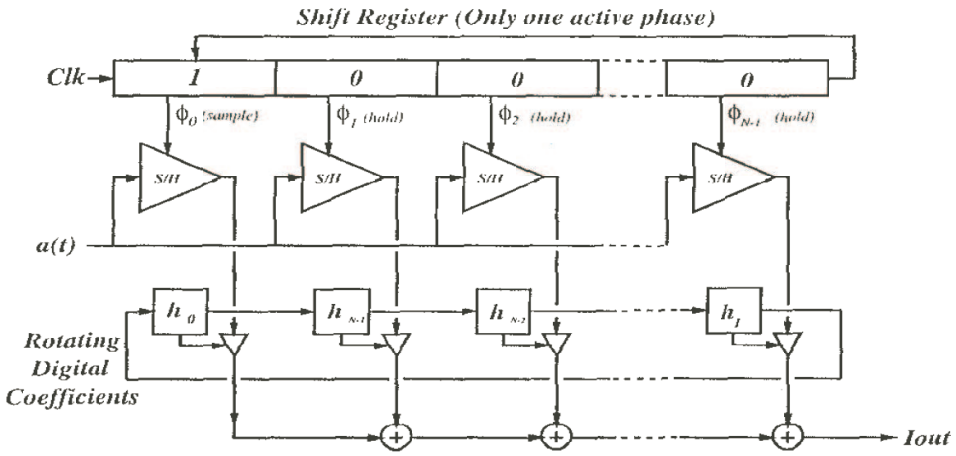


Figure 7. Example of switched-capacitor FIR filters. Sampled-data analog FIR filter with digitally programmable coefficients [44].

filter structure is usually employed [41]. Also, a parallel filter concept is suggested to increase the sample-rate-to-corner-frequency ratio of FIR filters [45]. In addition, a rotating switch matrix is used to eliminate the error accumulation [46]. Alternatively, these problems can be partially alleviated by employing over-sampling design techniques [42,47,48]. The filter implementations with these techniques offer a design flexibility by allowing for coefficient and/or input modulation [49, 50]. However, this design approach requires the use of higher clock rates to obtain high over-sampling ratios.

The programmability in analog FIR filter implementations can also be obtained by utilizing switched-current techniques. An example of switched-current FIR filter implementation is shown in Figure 7b. These techniques allow for the integration of the digital coefficients through the use of the current division technique [51] or multiplying digital-to-analog converters (MDAC) [52,53]. Moreover, a circular buffer architecture can be utilized to ease the problems associated with analog delay stages and to avoid the propagation of both offset voltage and noise [44, 54–56]. Recently, a switched-current FIR filter based on DA has also been suggested for pre-processing applications to decrease the hardware complexity and area requirements of the FIR filters [57].

1.5 Motivation for using floating-gate transistors in analog and mixed-signal circuits

In the previous sections, the overview of the techniques to deal with device imperfections and to obtain tunable and/or reconfigurable circuits is given. These techniques generally result in increase in power consumption and/or die area, which negate the benefits of the analog and mixed-signal circuits.

In this work, cooperative analog-digital signal processing (CADSP) approach is taken to design programmable circuits for signal-processing systems. In this respect, the design issues associated with the analog and mixed-signal circuits are circumvented by introducing floating-gate transistors to the available devices in the mainstream of the CMOS processes. This adopted approach, building tunable/reconfigrable circuits using floating-gate transistors, enables designers to exploit the benefits of the analog and mixed-signal circuits.

As illustrated in Figure 8, floating-gate transistors can be utilized in building tunable resistors and voltage references, which further extend the capabilities of the programmable circuit design. These circuit blocks are used in analog and mixed-signal circuit applications to demonstrate the tunability and reconfigurability as well as the low-power consumption and compactness. The tunable resistors are used in highly-linear amplifier and multiplier circuits to improve the linearity and to obtain precise resistors. Moreover, these resistors are utilized in building a binary-weighted resistor DAC. Similarly, tunable voltage references are employed in the implementation of a low-power and compact DAC and a reconfigurable distributed-arithmetic based FIR filter.

This thesis is organized into ten chapters. In Chapter 2, we present the design and the programming method of floating-gate transistors. In addition, we describe the necessary conditions for designing tunable resistors and the role of floating-gate transistors in these resistor implementations. After that we explain the design of a voltage reference and analyze its noise, temperature dependence, and charge retention. In Chapter 3, we describe the implementation of a tunable resistor using the gate-linearization technique and present

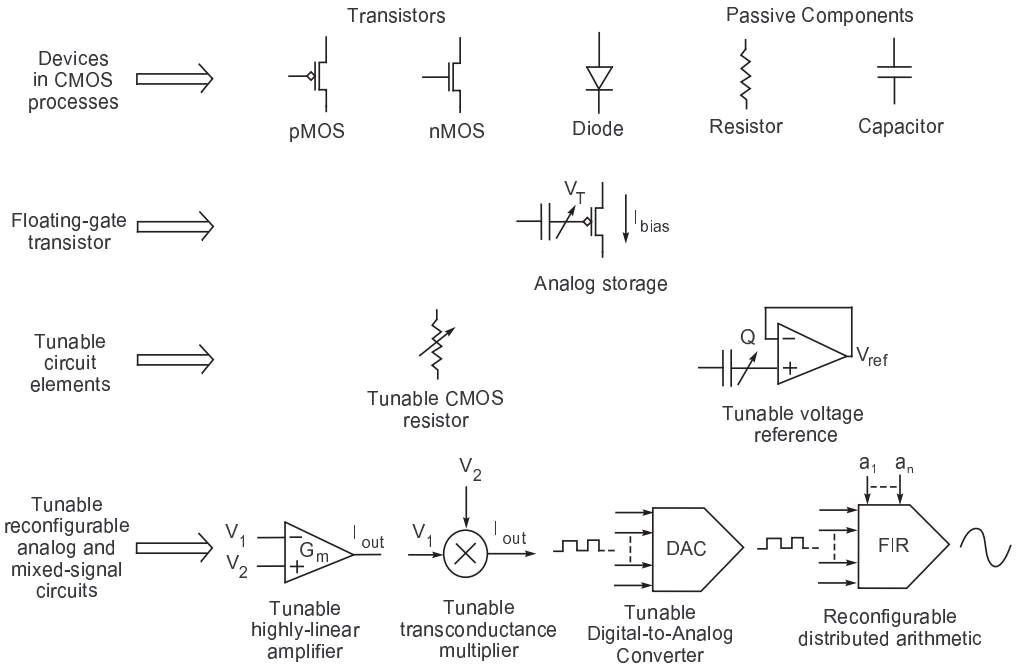


Figure 8. Design flow of the presented work. Floating-gate transistors are added to the available devices in the mainstream CMOS processes to design tunable resistive elements and voltage references, which are then used to build tunable and reconfigurable analog and mixed-signal circuits.

its experimental results. In Chapter 4, we explain the design and implementation of a compact tunable resistor using scaled-gate linearization technique and present its experimental results. In Chapter 5, we describe the implementation of a highly linear tunable resistor based on the common-mode linearization technique and compare it with other existing tunable resistors. In Chapter 6, we explain the design and implementation of a highly-linear amplifier and a transconductance multiplier employing the tunable resistor based on the common-mode linearization technique. In Chapter 7, we describe the implementation of a binary-weighted resistor DAC using the tunable resistor based on the scaled-gate linearization technique. In Chapter 8, we present the implementation of a programmable binary-weighted DAC using tunable voltage references and discuss the design issues. In Chapter 9, we describe the design and implementation of a reconfigurable distributed-arithmetic based FIR filter. Lastly, in Chapter 10, we discuss the impact of the presented work and describe the applications of the designed circuits and circuit blocks.

CHAPTER 2

DESIGN OF TUNABLE CIRCUITS USING FLOATING-GATE TRANSISTORS

The programmability of the floating-gate transistors enables to build systems that can adapt and/or be reconfigured. This allows to leverage the reconfigurability, which is generally associated with digital systems, into analog and mixed-signal circuits that are more area and power efficient. In this chapter, the tuning mechanisms of the floating-gate transistors are described. Also, the storage and capacitive coupling capabilities of floating-gate transistors to build tunable resistors and a voltage reference are also explained. These tunable resistors and voltage reference will be used to design and implement tunable and reconfigurable analog and mixed-signal circuits.

2.1 Floating-Gate Transistor Programming

The design of the floating-gate nMOS and pMOS transistors using regular nMOS and pMOS transistors is illustrated in Figure 9. Throughout this study, an indirect programming technique is utilized to tune the charge on the floating-gate terminal of these transistors. In this technique, a tunneling junction capacitor and an additional pMOS transistor are employed to tune the charge on the floating gate without introducing additional switches at the signal path.

Figure 10a illustrates that the threshold voltage of a floating-gate pMOS transistor can be increased or decreased by tuning the charge on the floating-gate terminal. The charge tuning is achieved by using hot-electron injection and Fowler-Nordheim tunneling mechanisms. The hot-electron injection increases the number of electrons on the floating-gate terminal; thus the threshold voltage of the pFET is decreased and the threshold voltage of the nFET is increased. In contrast, the tunneling mechanism decreases the number of electrons and has the opposite effect compared to hot-electron injection.

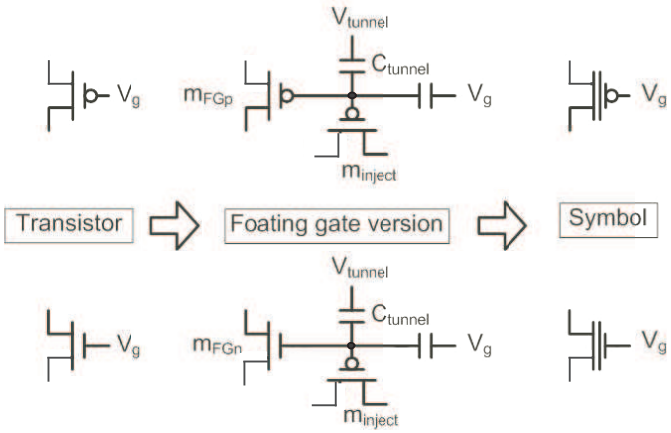


Figure 9. Design of floating-gate transistors from regular nMOS and pMOS transistors. Charge on the floating gate is tuned by employing Fowler-Nordheim tunneling and hot electron injection mechanisms. This is achieved by utilizing an indirect programming technique, where electrons are injected using a pMOS injection transistor, M_{inject}, and tunneled using a tunneling junction capacitor, C_{tun}.

The tunneling mechanism is utilized for the coarse programming of the threshold voltage. The rate of the electron tunneling can be increased by increasing V_{tun}. The precise programming, though, is done by employing the hot-electron injection mechanism. It is achieved by creating $6.5V$ voltage pulses across a pFET's drain and source terminals. These pulses are generated by modulating the drain terminal of M_{inject}, while keeping its source terminal fixed at $6.5V$.

Figure 10b illustrates that as the floating-gate voltage decreases, the injection efficiency drops exponentially since the injection transistor has better injection efficiency for smaller source-to-gate voltages. This efficiency drops as the transistor channel becomes more inverted. Therefore, the gate voltage, V_g, is modulated during programming to keep the floating-gate voltage at the same place, where the injection efficiency is high. In this way, the number of injected electrons and the output voltage change is accurately controlled. Moreover, it is observed that increasing the injection voltage, V_{sd}, increases the injection efficiency. However, after $6.5V$ the transistor channel becomes more inverted compared to the channel for a smaller injection voltage, and this degrades the injection efficiency.

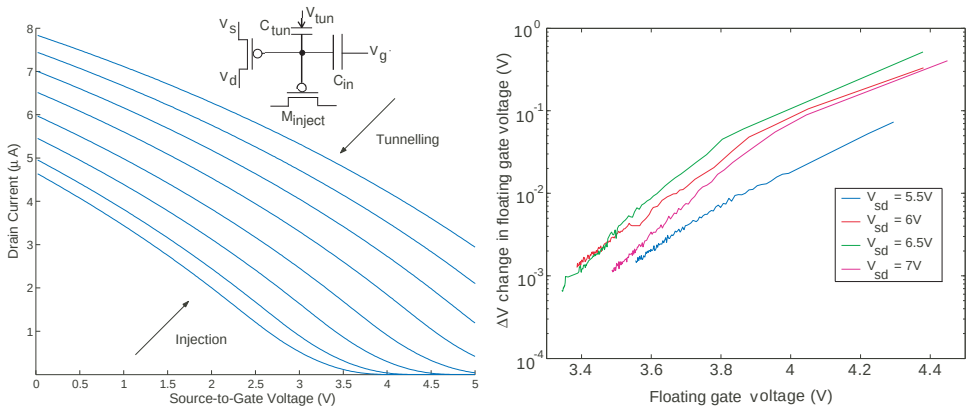


Figure 10. (a) Gate sweeps of a floating-gate pMOS transistor. The threshold voltage of the transistor is tuned by using Fowler-Nordheim tunneling and hot electron mechanisms. The threshold voltage can be made negative by increasing the number of electrons on floating gate using injection mechanism. (b) Change in the floating-gate voltage for 10ms injection pulses and for different injection voltages.

2.2 Tunable resistor design

The fundamental requirement to operate a single MOS transistor in the triode region as a linear resistive element is to suppress its nonlinearities by applying a function of the input signal to its gate [17] and/or its body [58]. In order to determine this function, the source of the nonlinearities in the drain current needs to be identified, and a linearization scheme has to be developed accordingly.

The drain current of a MOS transistor in the strong inversion has been accurately modelled [59], [60], [61]. Based on these models, three principal nonlinearities in the drain current of a long-channel transistor in the triode region are identified as the body effect, the mobility degradation, and the fundamental quadric component due to the common-mode of the drain and source voltages. These nonlinearities are mostly dependent on the common-mode of the input signals, and can be suppressed by building common-mode feedback structures around a transistor [18].

The linearization techniques based on the transistors operating in the triode region necessitate the generation of common-mode and large gate voltage for their proper operation. While most of the linearization techniques are appealing in terms of the reduced nonlinearity, building a feedback structure to generate a common-mode voltage generally results in

increased number of components and increased power consumption. In addition, creating a large quiescent voltage with fully integrated circuits in CMOS processes is not a trivial task. These disadvantages limit the operation of a linearized MOS transistor and, thus, the main tendency has been to look for alternative linearization techniques.

In this section, we show that introducing floating-gate MOS transistors can effectively circumvent these problems by providing capacitively coupled gate connection, and an quiescent gate voltage that can be adjusted by using the hot-electron injection and Fowler-Nordheim tunnelling mechanisms. The implementations of the linearization schemes will be described in the subsequent chapters.

2.2.1 Generation and tuning of a large quiescent voltage

For applications where tunable linear elements operating in triode region are required, creating a large DC offset voltage within the power supply voltage range becomes a crucial part of the design. This offset voltage is applied to the gate of the transistors to extend their triode operation regime. In this respect, a floating-gate transistor can be employed to alleviate this problem by generating a large offset that is not limited with the power supply.

The quiescent gate voltage ensures the proper operation of the linearized elements by keeping the transistors in the triode region, $V_{ds} < V_{gs} - V_T$, where V_{ds}, V_{gs}, and V_T are the drain-to-source, gate-to-source, and threshold voltages, respectively. The gate voltage is also utilized to control the resistance of these elements. Therefore, the operating range, which is determined by V_{ds}, has to be optimized to accommodate the desired tuning range of the resistor while still keeping the transistor in the triode region.

The drain sweeps of a regular pMOS transistor shown in Figure 11a illustrate that in a $0.5\mu m$ CMOS process, $V_{sg} > 5V$ needs to be supplied in order to keep the transistor in the triode region for $5V$ operating range. Although this allows to obtain the maximum linear operating range of the linearized elements, it necessitates the use of voltages that are larger than the power supply for nFETs or lower than the ground potential for pFETs.

When the common mode of the input signals is fixed, and a differential test is performed

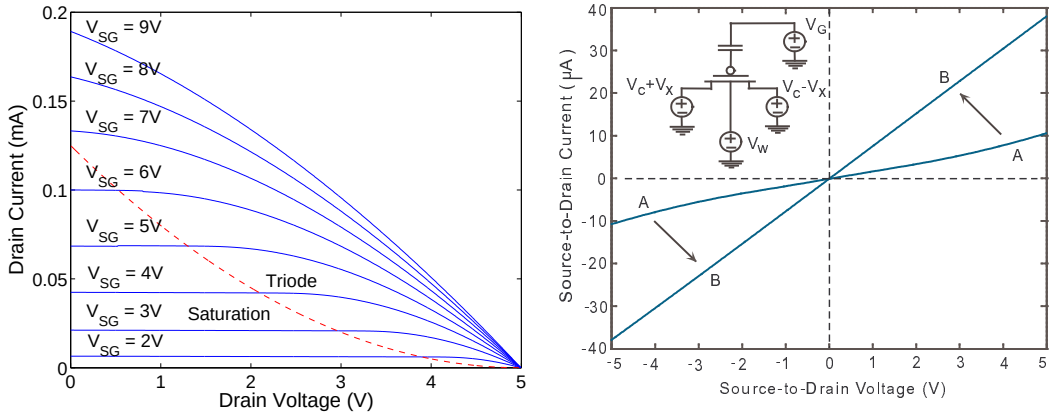


Figure 11. Drain sweeps of a pMOS transistor and differential test of a floating-gate transistor. (a) Drain voltage, V_d, sweep of a pMOS transistor tuned for gate voltages, V_G, from $2V$ to $9V$. Source and well voltages (V_s and V_w) are kept at $5V$. The dashed line separates the triode and saturation regions. (b) Differential test of a floating-gate CMOS transistor. The voltages, V_G, V_W, and V_C are set as $0V$, $5V$, and $2.5V$, respectively. V_X is swept from $-2.5V$ to $2.5V$. The curve-A is obtained without tuning the charge on the floating gate. The curve-B is measured after injecting electrons to the floating gate.

with a pMOS floating-gate transistor, as illustrated in Figure 11b, the drain current exhibits a linear characteristic as long as it stays in the triode region. The curve-A in Figure 11b is obtained without programming the floating-gate transistor. Assuming no extra charge is created on the floating gate, the output current of the floating-gate transistor for the differential test has the same characteristics as the output current of a regular MOS transistor with the same dimensions. In addition, it can be observed that for large drain-to-source voltages the transistor leaves the triode region, since $V_{ds} < V_{gs} - V_T$ does not hold anymore. However, after injecting enough electrons to the floating gate by using the hot-electron injection mechanism, the floating-gate voltage decreases much enough that the transistor exhibits a very linear characteristic for the given input voltage range. This is illustrated with the change in the transistor linearity from curve-A to curve-B in Figure 11b.

2.2.2 Common-mode voltage computation

The common-mode of the input signals can be computed using the capacitive design approach illustrated in Figure 12. This approach can readily allow for reduced power consumption without increasing the total harmonic distortion of the designed circuit. For input signals, V_1 and V_2, the capacitive division with capacitors, C_1 and C_2, results in an output voltage that can be expressed as

$$V_{out} = \frac{C_1}{C_1 + C_2} V_1 + \frac{C_2}{C_1 + C_2} V_2 + \frac{Q}{C_1 + C_2} \tag{1}$$

where Q is the charge stored at the capacitive node, V_{out}. If the capacitors are designed to be equal, the above expression becomes $V_{out} = (V_1 + V_2)/2 + V_Q$, where V_Q is the effect of the stored charge. Although the common-mode voltage can be computed precisely with this method, when the capacitors are used with a transistor, M, as shown in Figure 12, the input capacitance of the transistor cause error in the common-mode computation. In this case, for the same size input capacitors, $C_1 = C_2 = C$, the computed voltage becomes

$$V_{fg} = (V_1 + V_2)\frac{C}{2C + C_{in}} + V_s\frac{C_{gs}}{2C + C_{in}} + V_d\frac{C_{gd}}{2C + C_{in}} + V_b\frac{C_{gb}}{2C + C_{in}} + \frac{Q}{2C + C_{in}} \tag{2}$$

where C_{in} is the input capacitance of the transistor and composed of the gate-to-drain capacitor (C_{gd}), gate-to-source capacitor (C_{gs}), and gate-to-body capacitor (C_{gb}). Depending upon the transistor's region of operation, their values change with the input voltages. In the triode and saturation regions, C_{gb} becomes very small, thus can be ignored. In the triode region, C_{gs} can be expressed as

$$C_{gs} = \frac{2}{3} C_{ox} \frac{1 + 2\alpha}{(1 + \alpha)^2} \tag{3}$$

where $\alpha = 1 - V_{ds}(1 + \delta)/(V_{gs} - V_T)$, $\delta = \gamma/(2\sqrt{\phi_B + V_{sb}})$, and $C_{gd} = \alpha C_{gs}$ [61]. The crucial point here is that C_{gd} becomes equal to C_{gs} as $V_{gs} - V_T \gg V_{ds}(1 + \delta)$, which can be satisfied for deep threshold conditions. Moreover, in the saturation region of the transistor, C_{gd} becomes negligible, and C_{gs} becomes $2C_{ox}/3$.

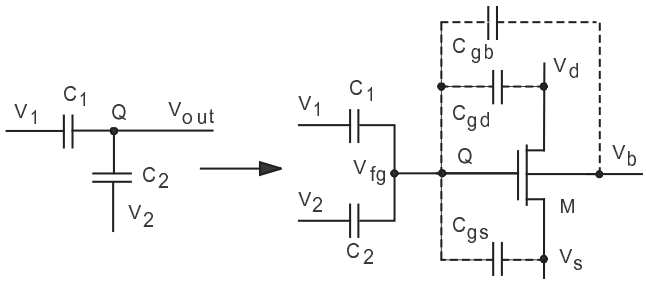


Figure 12. Common-mode voltage computation method using capacitive design strategy. The gate capacitors of an nMOS transistor are showed to illustrate their effect on the common-mode voltage computation when this transistor is integrated with input capacitors to form a floating gate.

These capacitor characteristics not only determine the limitations in implementing the linearization techniques, but also the amount of nonlinearity that can be suppressed with this approach. Therefore, the linearization techniques have to be built in consideration of the region of operation of the transistors and their capacitances.

2.3 Design of a tunable voltage reference

A tunable voltage reference can be built by using the analog storage feature of the floating-gate transistors. The design of a such voltage reference can enable to store the scaled-voltage levels for data converters and to obtain a tunability and reconfigurability for mixed-signal circuits. In this work, the tunable voltage reference (epot) is built to be incorporated into a voltage-output binary-weighted digital-to-analog converter (DAC) and a finite impulse response (FIR) filter.

Depending upon the application and its circuit specifications, the design of the epot can be different. For the DAC, the epot programming determines the programming precision and affects the maximum achievable DAC linearity. Also, the epot charge retention sets the lifetime of the DAC linearity. In addition, the temperature dependence of epots determines the operating range of the DAC, where the variation of the stored epot voltages with the temperature is less than the tolerable error. Similarly for FIR filters, the coefficients of the filters are stored by the epots, and thus the programming precision and charge retention of

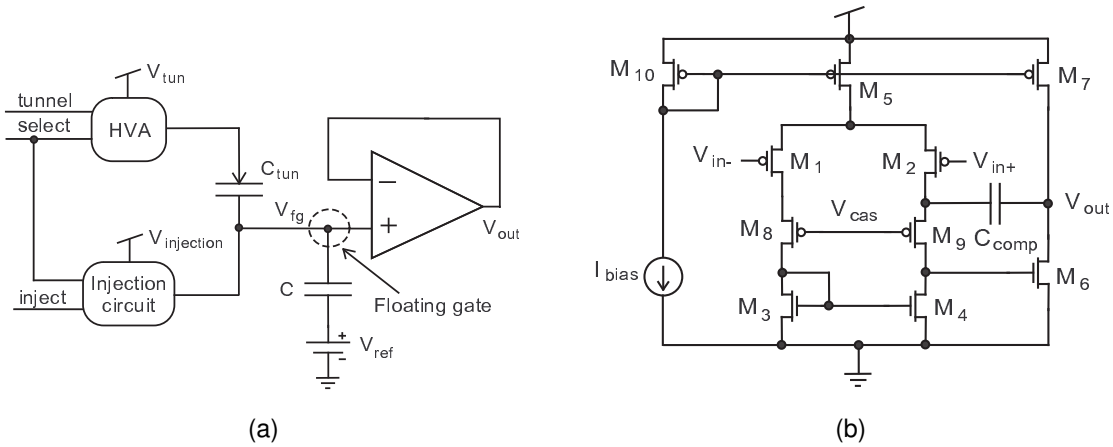


Figure 13. (a) Circuit schematic of the epot. Charge on the floating-gate is used to program the voltage output of the epot. The number of electrons on the floating gate are increased by using hot-electron injection and decreased by utilizing tunnelling quantum mechanical phenomena. The *tunnel*, *select*, and *inject* are the digital signals used for digital control of the epots. C_{tun} is the tunnelling junction used for tunnelling. (b) Low-noise amplifier used to buffer the stored voltage. V_{cas} is a bias voltage used for cascoding and C_{comp} is the compensation capacitor.

epots are also important design issues. Any change in the coefficients of the filter changes its characteristics and frequency response.

2.3.1 Epot programming

Epots are programmed using two methods defined as coarse and fine programming. The coarse programming is accomplished through the use of Fowler-Nordheim tunnelling, while the fine programming is performed by using hot-electron injection.

The epot programming circuitry is shown in Figure 14. In order to program the epots, the desired epot is first selected using a decoder and enabled by setting the *select* signal to high. Depending on whether the epot is to be programmed using the coarse or fine programming, the *digtunnel* or *digInject* signal is enabled, respectively. Programming of the epot involves the modification of the number of electrons on the floating node.

The tunnelling mechanism increases the epot voltage through the removal of electrons on the floating-gate node. The procedure for coarse programming of an epot involves tunnelling the epot until the epot output voltage reaches $200mV$ above the target voltage.

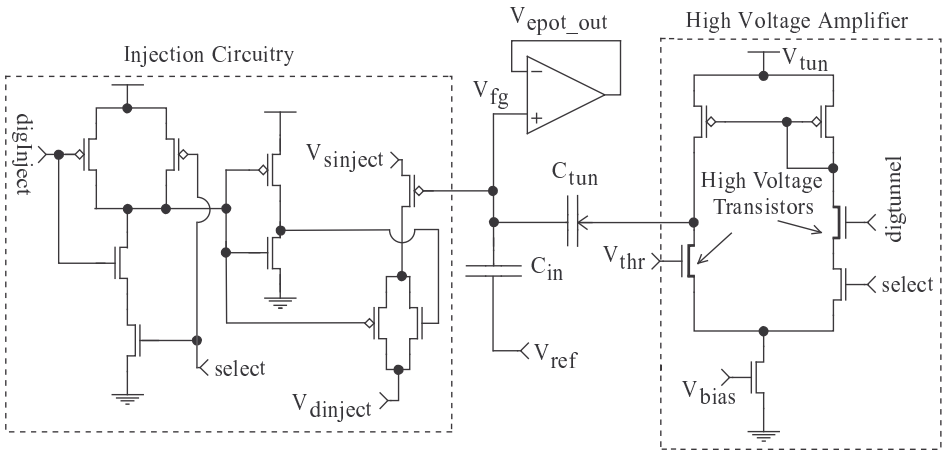


Figure 14. Programming circuitry of the epot. $V_{sinject}$ and $V_{dinject}$ are the source and drain voltages used to control the injection, while V_{tun} and V_{thr} are the tunnelling and the reference voltage of high voltage amplifier used to control tunnelling mechanism.

The purpose for overshooting is to avoid the coupling effect of the tunnelling junction on the floating-gate terminal once tunnelling is disabled. Once the *digtunnel* terminal is activated a high voltage is created across the tunneling junction. The high voltage amplifier is powered with $14V$ during coarse programming.

The hot-electron injection mechanism decreases the epot voltage through the addition of electrons onto the floating-gate node. Precise control of the injection process is achieved by pulsing $6.5V$ across the drain and source terminals of the pFET and by keeping the floating-gate voltage, V_{fg}, constant. By keeping V_{fg} constant, the number of injected electrons, hence the output voltage change, can be precisely controlled. During programming the input voltage, V_{ref}, of the epot is modulated based on the output voltage of the epot. This further facilitates precise programming since the epot output is approximately at the same potential as V_{fg}.

Once the output voltage of the epot has been programmed to the desired value through the use of coarse and fine programming, the tunnelling and injection voltages are set to ground to decrease power consumption, and to minimize the coupling to the floating-gate terminal.

2.3.2 Epot Noise

The data converters and mixed-signal circuits using epots to store their biases and references become sensitive to the epot noise. Also, when an array of epots are incorporated into VLSI circuits, these noise sources directly affect the linearity of the data converters and the characteristics of the circuits. The epot output noise can be written as

$$e_{epot}^2 = g_{m6}^2 R_{II}^2 \left[e_{n6}^2 + e_{n7}^2 + R_I^2 \left(g_{m1}^2 e_{n1}^2 + g_{m2}^2 e_{n2}^2 + g_{m3}^2 e_{n3}^2 + g_{m4}^2 e_{n4}^2 + \frac{e_{n8}^2}{r_{ds1}^2} + \frac{e_{n9}^2}{r_{ds2}^2} \right) \right] \tag{4}$$

where g_{m_i} is the transconductance of ith transistor, $R_I = r_{ds_{m4}} // (r_{ds_{m9}} + r_{ds_{m1}}(1 + r_{ds_{m9}} g_{m_9}))$, and $R_{II} = r_{ds_{m6}} // r_{ds_{m7}}$. Also, e_{ni}^2 can be written as

$$e_{ni}^2 = \left[\frac{8kT}{3g_{m_i}} + \frac{K}{fC_{ox}WL} \right] \tag{5}$$

In order to minimize the flicker noise, the amplifier is designed with pFET input stage. Also, input/load devices are sized properly to minimize the total epot output noise. The output noise of the epot is shown in Figure 15a. The epot voltage is measured through an on-chip buffer. Therefore, the measured epot noise also includes the noise of the buffer. The measured thermal noise level is $-120dB$, and the noise corner is measured to be around $4kHz$.

2.3.3 Epot temperature dependence

The temperature dependence of the epot is crucial for the circuits if the epot is employed to set their circuit parameters. The epot output voltage relative to the reference voltage can be written as

$$V_{epot} - V_{ref} = \frac{Q}{C} + V_{offset} \tag{6}$$

where V_{offset} is the offset voltage introduced by the epot amplifier. Assuming $\delta C/\delta T = \alpha$, where α is a process dependent parameter and around $50ppm/^oC$ for poly-poly capacitors [25], the temperature dependence of the relative epot voltage becomes

$$\frac{\delta(V_{epot} - V_{ref})}{\delta T} = -\alpha \cdot \frac{Q}{C} + \frac{\delta V_{offset}}{\delta T} \tag{7}$$

In addition, the temperature dependence of V_{offset} depends on the amplifier structure and the layout technique used to minimize the mismatch between the critical devices. In the proposed design, a common-centroid layout technique is employed to minimize the mismatch between input and load transistors, which are $M_1 - M_2$ and $M_3 - M_4$, respectively. This strategy helps minimizing the offset of the amplifier. If V_{offset} has temperature coefficient around $50ppm/^oC$, then it can be used to obtain a minimized temperature dependence. However, it is not possible to control this coefficient with the proposed design.

The epot output voltage for a range of temperatures is shown in Figure 15b, and the temperature coefficient is measured to be around $16.2ppm/^oC$. For an array of 10 epots programmed to different voltages, the mean temperature coefficient is measured as $16ppm/^oC$ with a maximum variation of $20.8ppm/^oC$. The epots are programmed relative to the reference voltage, which is set to $2.5V$.

2.3.4 Epot Charge Retention

After programming, it is crucial that the epots hold the stored charge for a long-term circuit reliability. The long-term charge loss of floating-gate transistors is mainly caused by the trap assisted tunnelling as well as the thermionic emission phenomenon [62,63]. By reducing the number of programming cycles, trap assisted tunnelling can be minimized. Since it may take the trapped electrons hours or days to be released from the traps, the initial programming is performed to come close to the desired epot voltage. After that minimized number of programming steps are applied for precise programming of the epots. The input capacitance of the epot can be sized properly to reduce the effect of the release of the trapped electrons.

Thermionic emission is a function of both temperature and time, and can be expressed as

$$\frac{V_{epot}(t)}{V_{epot}(0)} = \frac{Q(t)}{Q(0)} = exp\Big[- tv \cdot exp\Big(\frac{-\phi_B}{kT}\Big)\Big] \tag{8}$$

where $Q(0)$ and $Q(t)$ are the initial floating-gate charge and the floating-gate charge at time

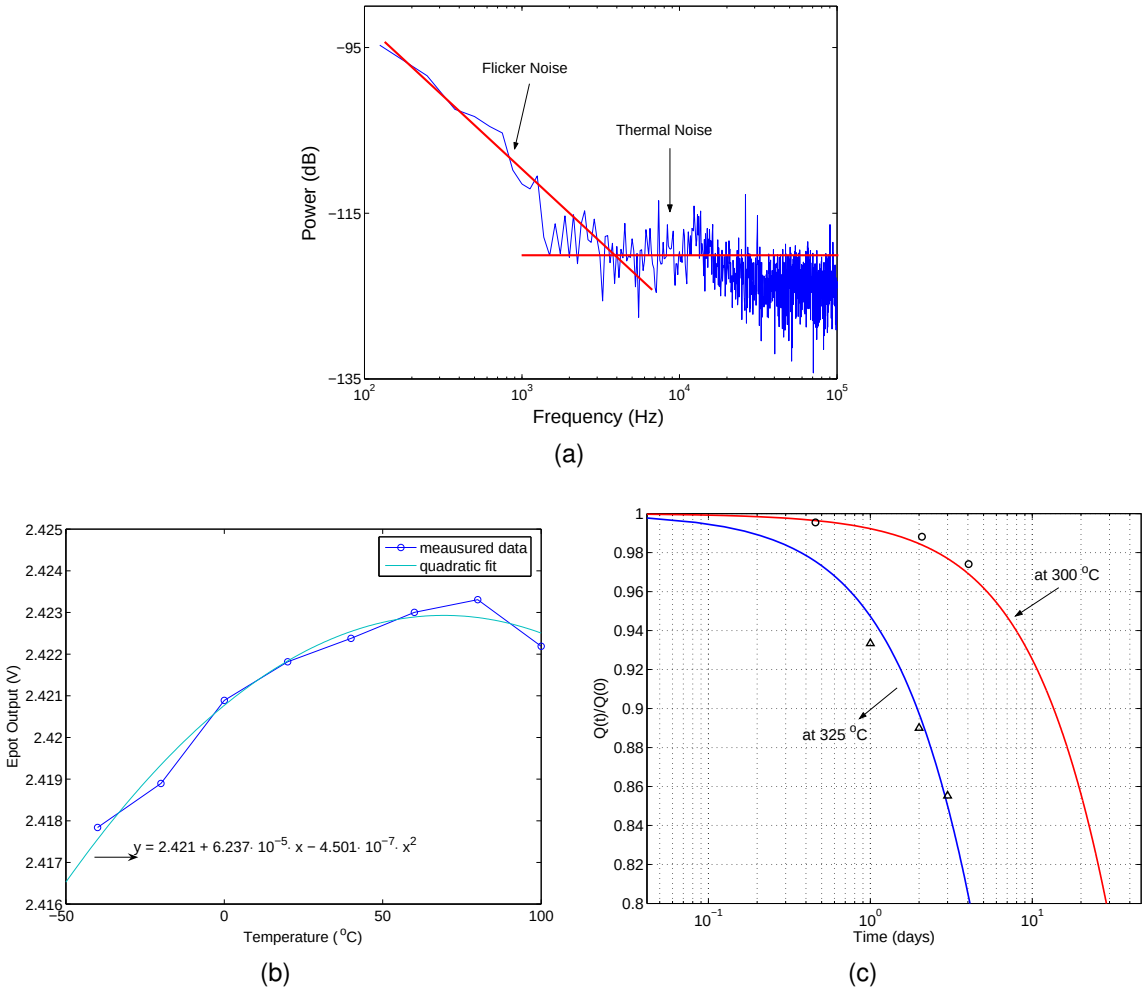


Figure 15. (a) Output noise of the epot. (b) Temperature sweep of the epot. The epot exhibits second-order temperature dependence when programmed around $2.422V$**.(c) Stress test performed at** 300^oC **and** 325^oC **to quantify the charge loss over time.**

t, respectively. Also, v is the relaxation frequency of electrons in poly-silicon, ϕ_B is the $Si - SiO_2$ barrier potential, k is the Boltzmann constant and T is the temperature. The change of the floating-gate charge directly affects the epot output voltage.

The design of the epots in CMOS processes with feature sizes smaller than $0.35\mu m$ necessitates the use of transistors with thicker silicon dioxide since the gate leakage becomes an serious issue in modern processes. Therefore, epots can be designed in these processes by using transistors with thick silicon-dioxide if they are available.

The retention of the epots are determined based on the stress tests. The theoretical fits using (8) along with the measurement results at 300^oC and 325^oC are shown in Figure 15c.

The worst case results are obtained after the first stress test at 300^oC. After the first test, the charge loss of the epots is decreased considerably. The ϕ_B and v from these worst-case experiments are extracted as $0.9eV$ and $60s^{-1}$. Based on this worst-case data, it is calculated that the stored epot voltage drifts $10^{-3}\%$ over the period of 10 years at 25^oC.

CHAPTER 3

A TUNABLE FLOATING CMOS RESISTOR USING GATE LINEARIZATION TECHNIQUE

Tunable CMOS resistors are are usually built based on the specifications imposed by their application. Depending upon the application, the CMOS resistors are generally required to be highly linear, area and power efficient, and to have a wide tuning/operating range. The compactness, power efficiency, and tuning range are the primary concerns for ANN systems. In this chapter, we present a tunable CMOS resistor that can be suitably employed in ANN systems. This CMOS resistor operates in the triode region, and utilizes the gate linearization technique [17]. In this structure, floating-gate transistors are not only employed to scale the input signals to the gate terminal [64], but also to store the charge on the floating gate to control the resistance.

In the next section, we explain the gate linearization strategy, and analyze its effect on the nonlinearities of a MOS transistor operating in the triode region. Subsequently, we describe the implementation of this technique to be used as tunable floating-gate resistor (FGR_{GL}). After that, we present the experimental results of this circuit. In the last part of the paper, we discuss its characteristics.

3.1 Gate Linearization Technique

For the circuits that are powered with a single power supply and required to operate rail-to-rail, it is generally the best design choice to fix the body/well potential of the linearized elements to one of the rails. In such cases, the gate linearization technique depicted in Figure 16 can be used to serve this purpose. This technique was originally proposed by Nay et al. [17], and used to suppress the fundamental quadratic component in the drain current. However, this technique does not completely eliminate the body effect and the mobility degradation.

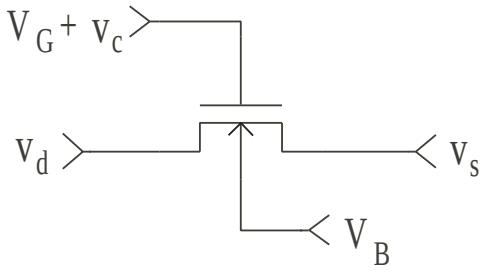


Figure 16. Gate linearization technique [17] applied to an nMOS transistor in the triode region. v_d and v_s are the drain and source voltages, respectively. V_G and V_B are the tunable quiescent gate and body voltages, and v_c is the common-mode voltage, $v_c = (v_d + v_s)/2$. The common-mode voltage is applied to the gate terminal to suppress the fundamental quadratic nonlinearity due to the common-mode of the drain-to-source input voltage.

The gate linearization is achieved by applying the common-mode signal, $v_c = (v_d + v_s)/2$, to the gate terminal with the addition of a tunable quiescent gate voltage, V_G. v_d and v_s are the drain and source voltages referenced to the ground, respectively. By using this technique, the quadratic term in the drain current is cancelled as shown in Appendix-I. In order for this technique to work effectively, the MOS transistor has to be kept in the triode region for the required input range. This requires $v_{ds} < 2(V_G - v_s - V_T)$, and also necessitates $v_g > V_T$, where V_T is the threshold of the device and can be expressed as

$$V_T = V_{FB} + \phi + \gamma\sqrt{v_c - v_b + \phi} \tag{9}$$

where V_{FB} is the flat-band voltage, ϕ is the surface potential, and γ is the body-effect coefficient. If θ_1, μ_1, V_{c_1}, and V_{G_1} are defined as

$$\theta_1 = \frac{\theta}{1 + \theta V_{G_1}}, \quad \mu_1 = \frac{\mu_0}{1 + \theta V_{G_1}} \tag{10}$$

$$V_{c_1} = \gamma\sqrt{v_c - v_b + \phi}, \quad V_{G_1} = (V_G - V_{FB} - \phi) \tag{11}$$

where μ_0 is the carrier mobility and θ is the mobility degradation factor, then, as shown in Appendix-I, the drain current for $\theta_1 \ll (V_{c_1} - \frac{v_{ds}^2}{96V_{c_1}^3})^{-1}$ can be approximated as

$$I_d = \frac{\mu_1 C_{ox} W}{L}\Big(v_{ds}(V_{G_1} - V_{c_1})(1 - \theta_1 V_{c_1}) + \frac{\gamma^4 v_{ds}^3}{96V_{c_1}^3}(1 + \theta_1(V_{G_1} - 2V_{c_1}))\Big) \tag{12}$$

where C_{ox} is the gate capacitance per unit area, W is the channel width, and L is the channel length. Ignoring the higher order terms, the resistance of the linearized element becomes

$$R = \frac{L}{\mu_1 C_{ox} W (V_{G_1} - V_{c_1})(1 - \theta_1 V_{c_1})} \tag{13}$$

By using the threshold equation in (9), and the θ_1 approximation, the resistance equation simplifies to

$$R = \frac{L}{\mu_1 C_{ox} W (V_G - V_T)} \tag{14}$$

This result reveals the fact that since V_T changes with v_c, the resistance of the linearized element depends on the common-mode of the input signals. In order to obtain higher linearity with this technique, it is necessary to have $V_G \gg V_T$.

The tunability with this structure can be obtained by by changing the value of V_G. Therefore, the tuning range of this resistor is limited by the required resistor linearity for the application and the maximum V_G that can be created.

3.2 Circuit Description

The gate linearization technique requires the generation of common-mode and large gate voltages for their proper operation. In this work, we show that introducing floating-gate MOS transistors provides capacitively coupled gate connection and an quiescent gate voltage that can be adjusted by using the hot-electron injection and Fowler-Nordheim tunnelling mechanisms. These features facilitate the circuit implementation of the gate linearization technique.

Employing a capacitive coupling connection to the gate terminal for the linearization was first suggested by Lande et al. [65], and implemented by using quasi floating-gate devices. However, a quasi floating-gate terminal acts as a high-pass filter with a very low corner frequency. Therefore, the DC common-mode of the input signals can not be tracked by the gate of the transistor with this approach. Here, we show that employing floating-gate transistors and using Fowler-Nordheim tunnelling and hot-electron injection quantum

mechanical phenomena for the resistance control improves the operation of CMOS resistors and their linearity.

The implementation of a tunable CMOS resistor based on the gate linearization technique is shown in Figure 17. This resistor operates as a floating resistor with its well terminal kept at a fixed potential. The common-mode voltage of the input signals is computed by using the feedback capacitors, which couple the drain and source voltages to the gate terminal. In addition, the charge stored on the floating-gate terminal creates the required quiescent gate voltage to satisfy the triode condition and linearity requirement. In this circuit, V_{tun} is used to enable the tunnelling mechanism to decrease the number of electrons on the floating-gate terminal. Also, V_{sPROG} and V_{dPROG} are used to create the required voltage difference that is necessary for the hot-electron injection mechanism to occur and increase the number of electrons at the gate terminal. As a result, the floating-gate voltage can be expressed as

$$V_{fg} = \frac{(C_g + C_{gs})V_s}{2C_g + C_{gs} + C_{gd} + C_{M_P} + C_{tun}} + \frac{(C_g + C_{gd})V_d}{2C_g + C_{gs} + C_{gd} + C_{M_P} + C_{tun}} + V_p \tag{15}$$

where V_p is the effect of the stored charge and the capacitive coupling from the peripheral circuit that includes C_{tun} and C_{M_P}. C_{tun} is the tunnelling junction capacitance, and C_{M_P} is the input capacitance of the injection transistor, M_P. In this equation, C_{gs} becomes equal to C_{gd} for large gate quiescent gate voltages. Therefore, the necessary condition for an accurate common-mode computation is to create a large quiescent gate voltage and to keep C_g much larger than C_{M_P} and C_{tun} so that the floating-gate potential is close to

$$V_{fg} \simeq \frac{(V_s + V_d)}{2} + V_p \tag{16}$$

The scaling error introduced by the common-mode computation increases the common-mode dependence of the circuit. However, the main source of the distortion in this linearization technique is the body effect since the body potential is fixed relative to the common-mode of the input signals.

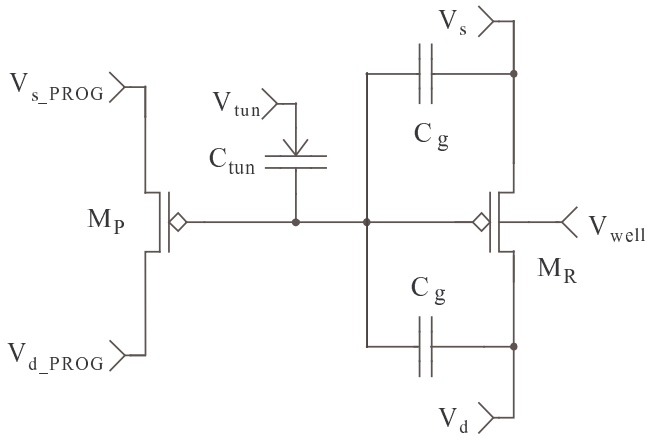


Figure 17. The circuit implementation of the gate linearization technique (FGR_{GL}). The common-mode feedback is realized by using feedback capacitors (C_g) between source-gate and drain-gate terminals. V_{well}, V_s and V_d are the well, source and drain voltages of M_R, respectively. This resistor is tuned by changing the quiescent gate voltage and this is achieved by using the tunnelling junction connected to V_{tun}, and the injection circuit that has source voltage V_{sPROG} and drain voltage V_{dPROG}.

3.3 Temperature dependence

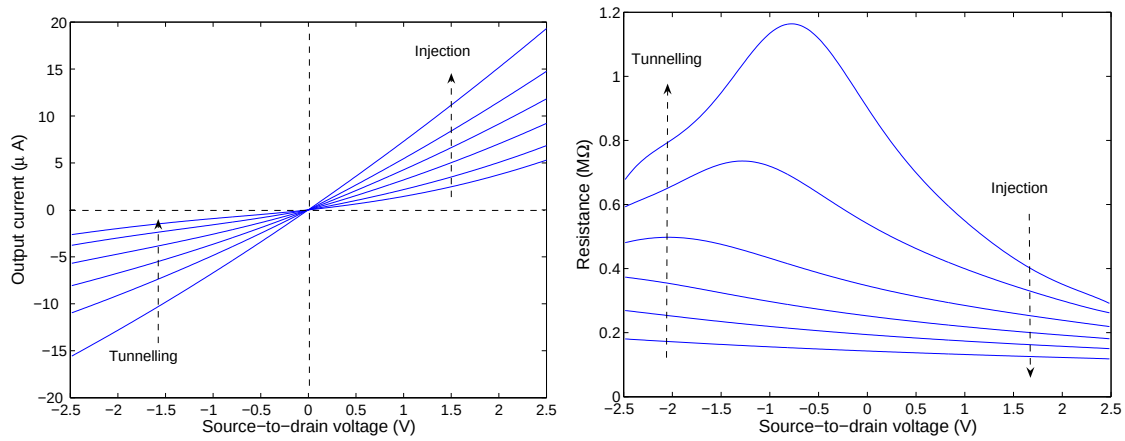
The temperature dependence of the FGR_{GL} can be found by ignoring the higher order terms in FGR_{GL} current and rearranging it as

$$\frac{1}{R} = \frac{\mu_n' C_{ox} W}{L}[V_G - V_T] \tag{17}$$

The temperature dependence of μ_n' and V_T can expressed as $\mu_n' = \mu_{n_o}'(T/T_0)^{-m}$ and $V_T = V_{T_0} - \alpha_{VT}(T - T_0)$, where T_0 is the reference temperature, and m is the positive constant that ranges from 1.5 to 2, and μ_{n_o}' and V_{T_0} are the temperature independent parameters. Also, α_{VT} is in the range of 0.5 to 4 $mV/^oC$ [61]. Hence, the temperature coefficient of the FGR_{GL} can be expressed as

$$\frac{1}{R}\frac{\delta R}{\delta T} = -\frac{1}{\mu_n'}\frac{\delta \mu_n'}{\delta T} + \frac{1}{V_G - V_T}\frac{\delta V_T}{\delta T} = \frac{m}{T} - \frac{\alpha_{VT}}{V_G - V_T} \tag{18}$$

where $\frac{1}{\mu_n'}\frac{\delta \mu_n'}{\delta T} = -\frac{m}{T}$ and $\frac{\delta V_T}{\delta T} = -\alpha_{VT}$. As a result, the temperature coefficient of the FGR_{GL} can be tuned by altering the effect of α_{VT} through the use of V_G. For desired temperature, T_d, and $V_G = \frac{\alpha_{VT}}{m}T_d + V_T$, the temperature coefficient of the FGR_{GL} can be set to zero at T_d.



(a) (b)

Figure 18. Experimental results. (a) I-V characteristics of the FGR_{GL}. (b) Extracted resistances of the FGR_{GL} tuned to different quiescent gate voltages.

3.4 Experimental results

In this section, we present the characterization results of the proposed circuit. The measurements were obtained from the chip that was fabricated in a 0.5μm CMOS process.

The experiments for the static measurements are performed by keeping one terminal of the floating-gate resistors at $2.5V$, and then sweeping the other terminal between 0 and $5V$. Also, the well terminal of FGR_{GL} is kept at $5V$. After each programming step by tuning the quiescent gate voltage, the experiment is repeated to observe the change in the resistance and linearity. The I-V curves of the FGR_{GL} are shown in Figure 18a. FGR_{GL} exhibits better linearity for its smaller resistance values. This is mainly because the relative effect of the common-mode voltage on the resistance becomes less for higher V_G voltages. The extracted resistance sweeps of the FGR_{GL} are shown in Figure 18b. It can be observed that the resistance of FGR_{GL} changes with the input voltage. This agrees with the theoretical results shown in (13), since V_T deviates from its nominal value with the change in the common-mode voltage. Therefore, this resistor exhibit small changes for small resistance values, and large changes for large resistance values. While decreasing the resistance of the floating-gate resistor the quiescent gate voltage is also increased. In turn, this helps the transistor to stay in the deep triode region even for large differential input signals.

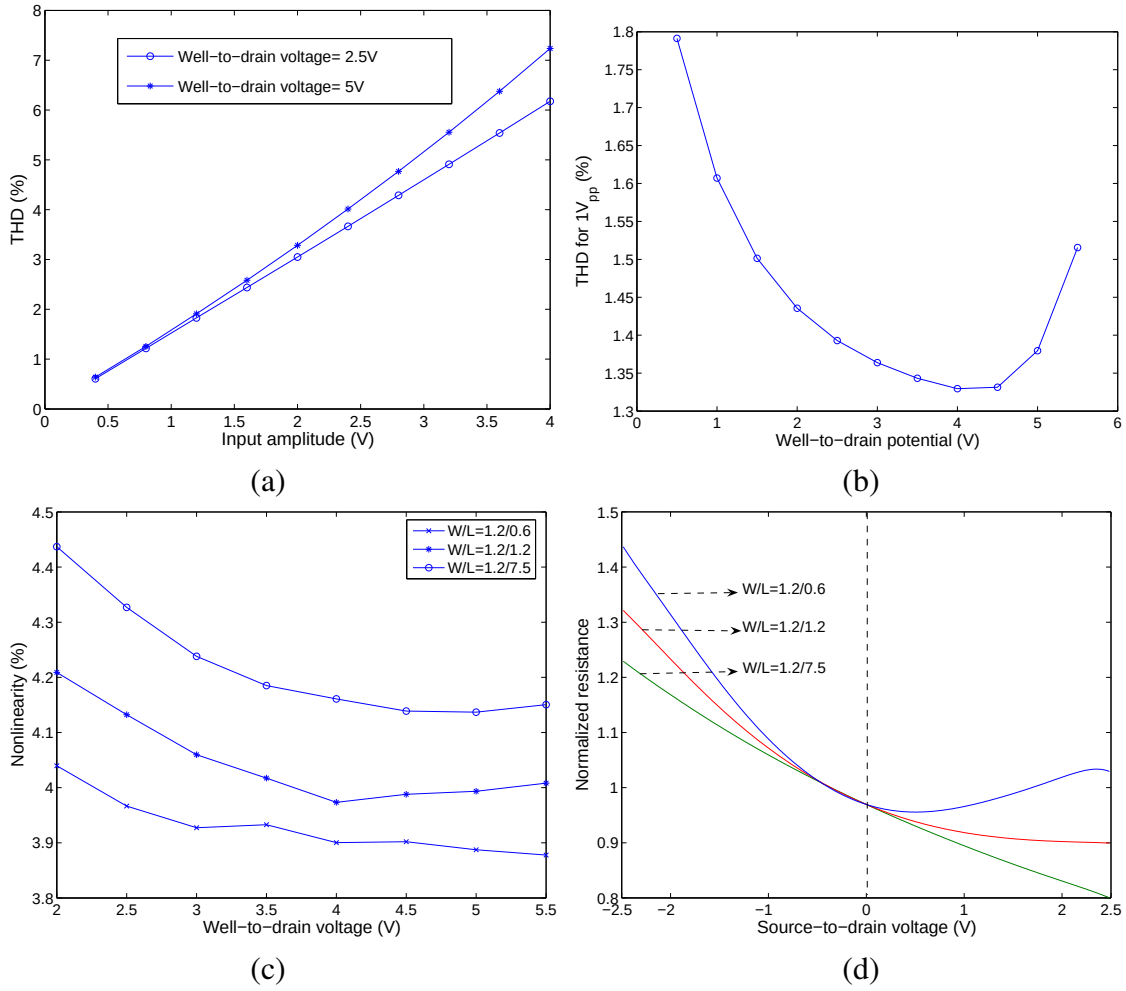


Figure 19. Experimental results. (a) Total harmonic distortions of the FGR_{GL} for a range of sine-wave signal amplitudes. (b) Total harmonic distortions of the FGR_{GL} for $1V_{pp}$ sine wave signal and for a range of well voltages. (c) Total nonlinearity of the FGR_{GL} circuits in the full range of operation (0-5V). The length of the transistors are $0.6\mu m$, $1.2\mu m$, and $7.5\mu m$. (d) Normalized resistances of the FGR_{GL} circuits for different transistor lengths.

The dynamic measurements of floating-gate resistors are obtained by using an off-chip inverting amplifier with a corresponding feedback resistor (matches the resistance of on-chip resistor). Also, 16 – *bit* DAC is employed to generate the sine-wave for the characterization of the FGR_{GL} linearity. The distortion level of this resistor for a range of signal amplitudes is illustrated in Figure 19a. This experiment is repeated for $V_{well} = 5V$ and $V_{well} = 7.5V$ while V_{drain} is kept at $2.5V$ and V_{source} is swept around $2.5V$. It is observed that the linearity is also dependent on the well-to-drain potential. Therefore, another linearity test is performed for a range of well-to-drain voltages as shown in Figure 19b. For

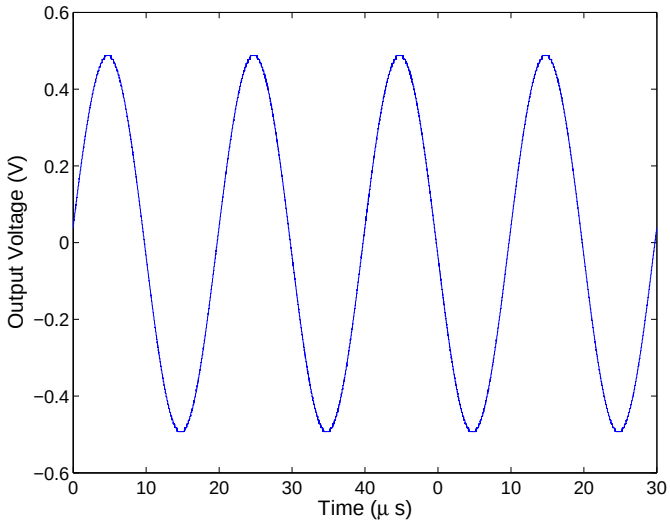


Figure 20. Experimental results. Transient response of the FGR_{GL} for $1V_{pp}$ $100kHz$ sine-wave.

$1V_{pp}$ sine-wave, it is seen that the linearity of the FGR_{GL} can be increased by keeping the well-to-drain voltage around $4V$. The main source of the distortion in FGR_{GL} linearity is the change in its threshold due to body effect, and this effect becomes more apparent as the resistance of FGR_{GL} is increased by decreasing the quiescent gate voltage. Depending upon the linearity level that certain applications may require the tuning range of these resistors can be determined.

The change in the total nonlinearity of the FGR_{GL} for different transistor lengths is depicted in Figure 19c. The transistor lengths are chosen as $0.6\mu m$, $1.2\mu m$, and $7.5\mu m$. It is observed that although the second-order effect in short-channel transistors becomes more dominant, the total nonlinearity in the full range of operation becomes less for short-channel transistors. This is mainly due to their resistance behavior with the common-mode voltage. As shown in Figure 19c, transistors with shorter channels exhibit less variation with the input and common-mode voltage change. Moreover, the transient test of the FGR_{GL} is performed by using $1V_{pp}$ $100kHz$ input signal. It is seen that FGR_{GL} operates at $100kHz$ without being limited by its feedback capacitors.

The temperature test of the FGR_{GL} is performed to characterize its temperature dependence between -60 to $100\ ^{o}C$. As shown in Figure 21a, the temperature behavior of the

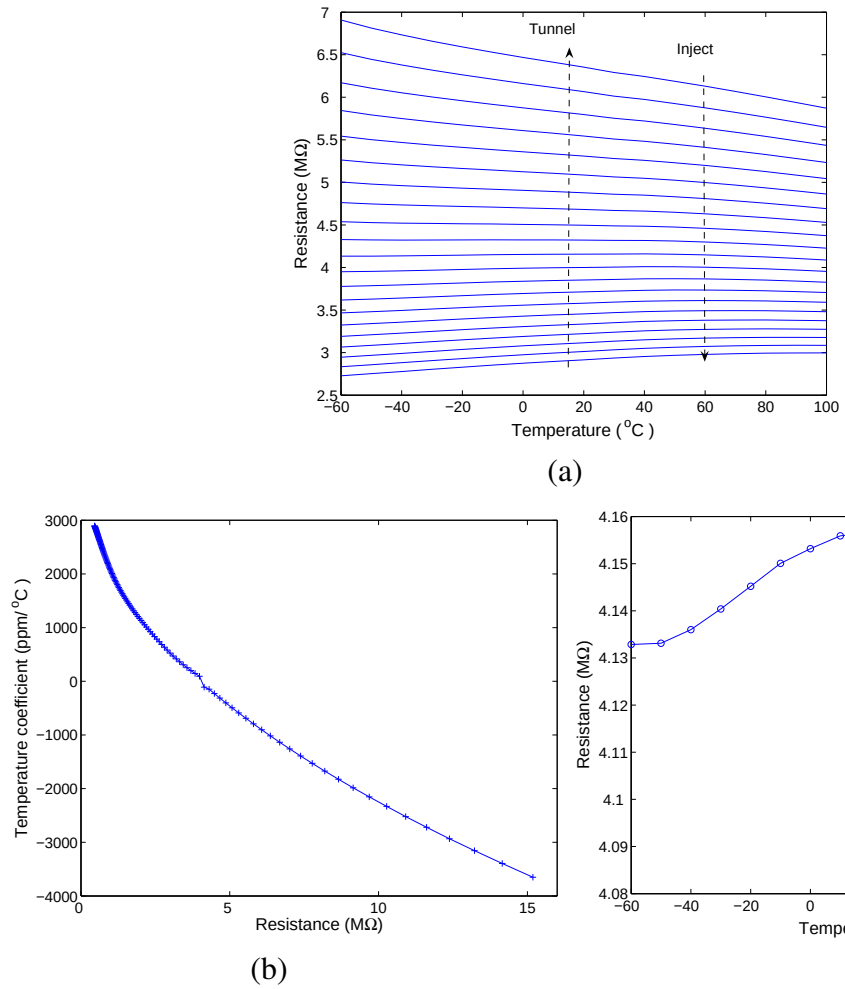


Figure 21. Experimental results. (a) Temperature behavior of FGR_{GL} for differently tuned resistance values. (b) Temperature coefficient of the FGR_{GL} for a range resistance values. (c) Temperature behavior of the FGR_{GL} when its first-order temperature dependence is cancelled.

FGR_{GL} depends on the programmed resistance value. Figure 21b illustrates the change of the temperature coefficient of the FGR_{GL} with the programmed resistance value. Around $10M\Omega$, the first-order temperature dependence of the FGR_{GL} becomes much less than its second-order temperature dependence, thus for this operating condition FGR_{GL} is governed by its second and higher-order temperature dependence. As shown in Figure 21c, the temperature coefficient of the FGR_{GL} can be reduced down to $106ppm/^oC$.

Finally, the die photo of the fabricated FGR_{GL} circuit is shown in Figure 22. The total area of this circuit is $4900\mu m^2$ and its each gate-feedback capacitor is $1.46pF$.

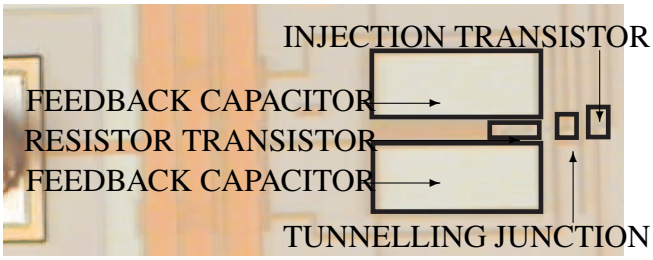


Figure 22. Die photo of the fabricated FGR_{GL} circuit.

3.5 Discussion

The presented CMOS resistor is very suitable for variety of applications. Especially alleviating the trade-off between the tuning range and the operating range by using floating-gate transistors allows to leverage the tunability into analog circuits without being limited by the supply rails.

The floating-gate resistor reported in this chapter utilizes the properties of MOS transistors in a CMOS process. FGR_{GL} uses only 2 capacitors and 1 transistor in addition to the programming circuit. It yields around 1.3% linearity (for $1V_{pp}$) without consuming additional power for the operation of the circuit. Moreover, FGR_{GL} can be easily employed in low-voltage applications since the operation of FGR_{GL} does not depend on any of the supply rails. Therefore, FGR_{GL} offers a circuit implementation of a power efficient, compact, and tunable CMOS resistor. Especially, this design becomes very suitable for the ANN systems, where an array of compact CMOS resistors needs to be integrated while keeping the power consumption down. Finally, this resistor has the ability to store its own resistance value, therefore it does not require an additional circuit to generate a voltage to set its resistance.

CHAPTER 4

A TUNABLE FLOATING-GATE CMOS RESISTOR USING SCALED-GATE LINEARIZATION TECHNIQUE

The tunable CMOS resistors offer a design flexibility in building precision and compact analog circuits. Therefore, they are widely used in transconductance multipliers, highly linear amplifiers, and tunable MOSFET-C filters. While the passive resistors that are implemented by using polysilicon, diffusion, or well strips in a CMOS technology exhibit around ±0.1% matching accuracy and around ±30% tolerance [33], the tunable CMOS resistors easily achieve high and precise resistance values through the utilization of controllable MOS channel resistance.

ANN systems as well as other low-power and low-voltage applications require a design of a compact and tunable resistor that is not only less sensitive to the mismatches but also suitable for the operation at low supply voltages. Moreover, such a resistor needs to achieve the required linearity and tuning range with low power consumption. In this chapter, we propose such a tunable CMOS resistor that can be successfully incorporated into ANN systems as well as low-power and low-voltage applications. This CMOS resistor employs a floating-gate transistor operating in the triode region, and utilizes the scaled-gate linearization technique [18] to decrease its nonlinearities. In the next section, we explain the scaled-gate linearization technique, and theoretically analyze the nonlinearities of a MOS transistor operating in the triode region. Subsequently, we describe the circuit implementation of this tunable floating-gate resistor (FGR_{SGL}). In the last part of this chapter, we present the experimental results of this resistor.

4.1 Scaled-gate linearization technique

In a standard CMOS technology, a linearized tunable CMOS resistor can be designed by employing a linearization technique. Such techniques exploit MOSFET's square-law characteristic in the saturation region [7], [8] or its resistive nature in the triode region [9]. In the triode region, the common-mode [18], gate [17], and scaled-gate linearization [18] techniques can be easily utilized to suppress the nonlinearities of a MOS transistor and to design a tunable CMOS resistor.

In contrast to other strategies, the common-mode strategy offers a high linearity, but its implementation requires the use of a higher voltage than the supply voltage and increased resistor area to generate the well-feedback voltage [66]. If high linearity is traded with a simplified design to suppress only the fundamental quadratic component of transistor nonlinearity, then the gate linearization technique can be utilized to build a compact tunable resistor [67]. Alternative to these techniques, the scaled-gate linearization technique can be adopted to a single MOS transistor in the triode region to alleviate the area and linearity issues of the tunable CMOS resistors.

The scaled-gate linearization scheme is depicted in Figure 23 and realized by applying a scaled common-mode voltage to the gate terminal. If the body potential, v_b, is fixed at some bias potential, V_B, then the fundamental quadratic component, $v_{ds} \cdot v_c$, and the body effect can be cancelled by applying a scaled common-mode voltage to the gate terminal. This is achieved by choosing the scale factor as [18]

$$a = 1 + \frac{\gamma}{2\sqrt{(V_B + \phi)}} \tag{19}$$

where γ is the body-effect coefficient and ϕ is the surface potential.

For a fixed body potential, V_B, a becomes a process dependent parameter. While the variation in process parameters becomes a limiting factor for this technique, this can be overcome by tuning V_B. After applying this technique, the first order mobility dependence

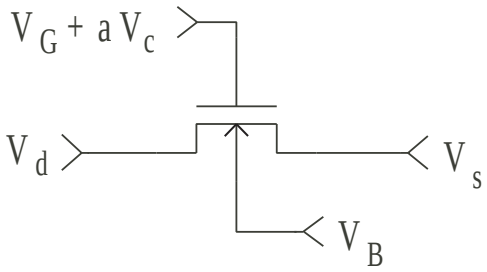


Figure 23. Scaled-gate linearization technique [18] to eliminate the fundamental quadratic nonlinearity and body-effect of an nMOS transistor operating in the triode region. The scale factor a is a process and body voltage dependent parameter. V_d and V_s are the drain and source voltages, respectively. V_G and V_B are the tunable quiescent gate and body voltages, and V_c is the common-mode voltage, $(V_d + V_s)/2$.

of the transistor dominates the distortion, and the drain current can be approximated as

$$I_d = \frac{\mu_o' C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} - \frac{\theta'\gamma[V_G - V_T](v_{ds}v_c)}{\sqrt{(V_B + \phi)}}\right\} \tag{20}$$

where C_{ox} is the gate capacitance per unit area, W is the channel width, L is the channel length, V_G is the quiescent gate voltage, and V_T is the threshold voltage. Also, v_{ds} is the drain-to-source voltage, v_c is the common-mode voltage and equal to $(v_d + v_s)/2$, and μ_o' and θ' are

$$\mu_o' = \frac{\mu_o}{1 + \theta[V_G - V_{FB} - \phi + \gamma\sqrt{(V_B + \phi)}]} \tag{21}$$

$$\theta' = \frac{\theta}{1 + \theta[V_G - V_{FB} - \phi + \gamma\sqrt{(V_B + \phi)}]} \tag{22}$$

where V_{FB} is the flat-band voltage, μ_0 is the carrier mobility, θ is the mobility degradation factor.

This technique is used to eliminate not only the fundamental quadratic term, but also the body effect term. The input voltage range of this technique is determined by the triode condition, which is $V_{ds} < \frac{2}{(2-a)}(V_G - V_T - (1-a)V_s)$ for $V_g = V_G + a(V_d + V_s)/2$. Therefore, this technique requires the design of a scale factor to minimize the nonlinearities, and the generation of a large V_G to ensure the triode operation for given operating range.

4.2 Circuit Description

The circuit implementation of the FGR_{SGL} is shown in Figure 24. In contrast to the floating-gate implementations of the gate linearization [67] and common-mode linearization [66] techniques, one of the input terminal of FGR_{SGL} has to be maintained at a fixed potential, or at AC ground. Use of a floating-gate transistor in this structure enables to obtain the scale factor and large gate voltages due to its capacitive coupling and charge storage capabilities.

The scale factor, a, is obtained by sizing the transistors and capacitors connected to the floating-gate terminal. Since the common-mode voltage and the scale factor in this structure are computed at the same time, the scale factor for this implementation can be redefined as $\chi = a/2$. With this implementation, χ can be expressed as

$$\chi = \frac{C_{g_1}}{C_{g_1} + C_{g_2} + C_P + C_{M_R}} \tag{23}$$

where C_{g_1} and C_{g_2} are the gate feedback capacitor and the trimming capacitor. Also, C_P, and C_{M_R} are the parasitic capacitance of the peripheral circuit and input capacitance of M_R, respectively. C_{M_R} consists of the gate-to-drain capacitor (C_{gd}), gate-to-source capacitor (C_{gs}), and gate-to-well capacitor (C_{gw}). In triode region, $C_{gs} = \alpha C_{ds}$, where $\alpha = 1 - V_{ds}(1 + \delta)/(V_{gs} - V_T)$ and $\delta = \gamma/(2\sqrt{\phi_B + V_{sb}})$ [61]. Since a part of C_{M_R} contributes to C_{g_1}, this effect needs to be taken into account when designing the circuit with large transistors.

Moreover, the charge on the floating-gate terminal is tuned by employing an indirect programming scheme. In this scheme, a tunnelling junction capacitor and an additional $pMOS$ transistor are used to tune the charge on the floating-gate terminal without introducing additional switches in the signal path. The resistance of FGR_{SGL} is tuned by utilizing the Fowler-Nordheim tunnelling and hot-electron injection quantum mechanical phenomena. V_{tun} is used to enable the tunnelling mechanism to decrease the number of electrons; and V_{sPROG} and V_{dPROG} are used to create the required voltage difference that is necessary for the hot-electron injection mechanism to occur and increase the number of electrons on the floating-gate terminal.

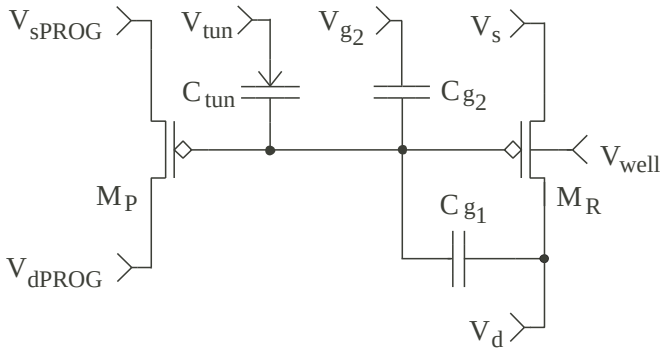


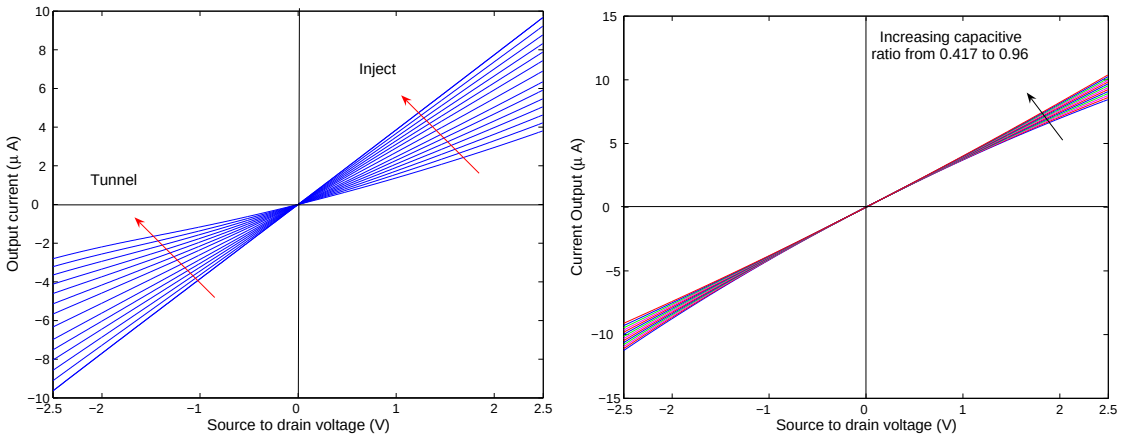
Figure 24. The circuit implementation of the scaled-gate linearization scheme (FGR_{SGL}). V_{well}, V_s and V_d are the well, source and drain voltages of M_R, respectively. Also, V_{sPROG} and V_{dPROG} are the source and drain voltages of the injection transistor, M_P. V_s is kept constant at ac ground, and V_d is used as the input of the resistor. V_{g_2} is held fixed during normal operation. The resistance tuning is achieved by changing V_{g_2} or the charge on the floating gate. The floating-gate charge is tuned by using the tunnelling junction, C_{tun}, connected to V_{tun} for Fowler-Nordheim tunnelling, and employing M_P for hot electron injection.

In addition to the programming circuit, this structure has only one transistor and two capacitors resulting in a very compact circuit. The scale factor has to be chosen properly to minimize the nonlinearities. However, there is no specific matching between the devices necessary. Therefore, the total area can easily be optimized for given application. Furthermore, since the computation is achieved by utilizing the capacitive coupling and charge storage capabilities of the floating-gate transistors, no additional power consumption is needed. This feature is especially useful for low-power applications.

4.3 Experimental results

In this section, we present the characterization results of the proposed circuit and discuss its features based on the measurement results. The measurements are obtained from a chip that was fabricated in a 0.5μm CMOS process. The test structure is built to have one main capacitor and thirteen trimming capacitors that are used to change the scaling factor.

The static measurements of the FGR_{SGL} are shown in Figure 25a and obtained by keeping the source terminal at $2.5V$, and then sweeping the drain terminal from $0V$ to $5V$. The scale factor, χ, for this experiment is chosen as 0.7918. Also, the well terminal of the circuit is kept at $5V$. After each programming step using tunnelling and injection, the experiment

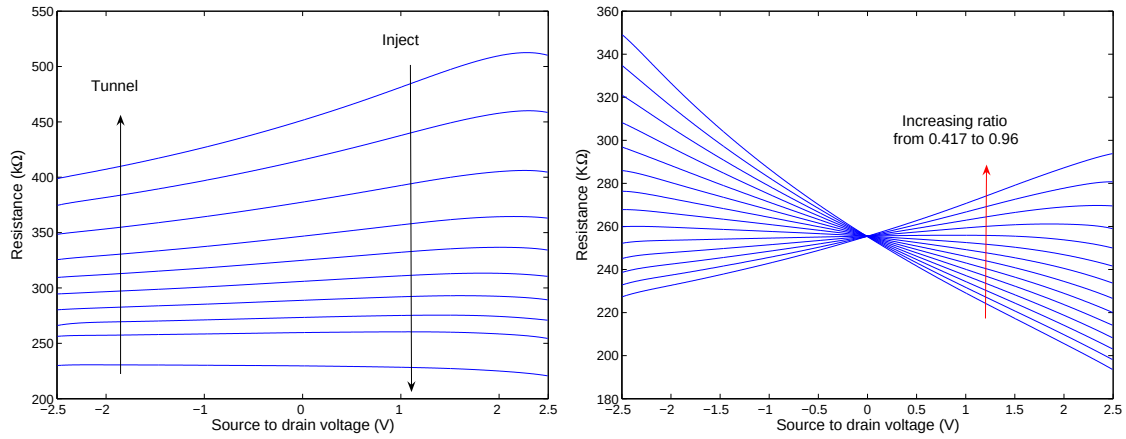


(a) (b)

Figure 25. (a) Experimental results obtained with differently programmed quiescent gate voltage (V_G) for $\chi = 0.7918$. V_G is increased through injection to decrease the resistance. Similarly, V_G is decreased by using tunnelling to increase the resistance. (b) The effect of χ on the linearity of the resistor. χ is increased from 0.417 to 0.96, and the implemented scale factors are 0.417, 0.4584, 0.5001, 0.5418, 0.5835, 0.6251, 0.6668, 0.7085, 0.7502, 0.7918, 0.8335, 0.8752, 0.9169, and 0.96.

is repeated to observe the change in the resistance and linearity. Figure 25b illustrates the effect of the scale factor on the linearity and resistance of the FGR_{SGL}. It is observed that the second-order nonlinearity is compensated well especially when the scale factor is chosen as 0.7918. The second-order nonlinearity is more apparent for scale factors smaller than 0.7918. As the scale factor is increased up to 0.96, the nonlinearity is increased too. Therefore, the optimum value for the scale factor is found as 0.7918.

The extracted resistance of the FGR_{SGL} for differently tuned resistance values are shown in Figure 26a. The scale factor is fixed at 0.7918, and the resistance is again changed by using the tunnelling and injection mechanisms. It is observed that as more electrons are injected to the floating gate, which means increasing V_G for a pMOS transistor, FGR_{SGL} becomes more linear. This is mainly because the triode condition for the resistor is satisfied more with the increased V_G values. As the resistor operates in deep triode region, it allows for larger voltage swings across its terminals. In addition, the relative nonlinearity of the transistor decreases for higher V_G values since θ' in (22) reduces for higher V_G values. The extracted resistances for different scale factors are shown in Figure 26b. These sweeps



(a) (b)

Figure 26. Experimental results. (a) Extracted resistances of the FGR_{SGL} tuned to different quiescent gate voltage. (b) Extracted resistance of the FGR_{SGL} for different scale factors, from 0.417 to 0.96, used to linearize the resistor. The sweeps show the voltage dependence of the resistor, and illustrates the compensation of the second order nonlinearity.

justify the previous result that the optimum value of the scale factor for better linearity is 0.7918. Also, the extracted resistances that have a smaller or larger scale factor than 0.7918 exhibit a non-symmetric behavior due to the fact that the second-order nonlinearity becomes the dominant source of the nonlinearity if it is not cancelled properly. This non-symmetric behavior is also partially contributed by the fact that C_{gd} and C_{gs} of M_R have unequal voltage dependent values.

The low-voltage characteristics of FGR_{SGL} are determined by decreasing the well voltage down to $0.25V$, as illustrated in Figure 27a. Each sweep in this plot is performed by changing V_d from $-V_{well}$ to $+V_{well}$ while keeping V_s at $V_{well}/2$. The sweeps are obtained for V_{well} equal to $0.25V$, $0.5V$, $1V$, $2V$, and $4V$. It is observed that the linearity of the resistor is preserved at low-voltages even if the capacitive division factor to obtain the scale factor is fixed at 0.7918. Since the well voltage changes the effective scale factor, the circuit should be designed for the desired supply voltage that also determines the well voltage. However, the results of this test show that the change in the effective scale factor due to the well voltage does not alter the linearity characteristics of the resistor as much as the change due to the capacitive division factor.

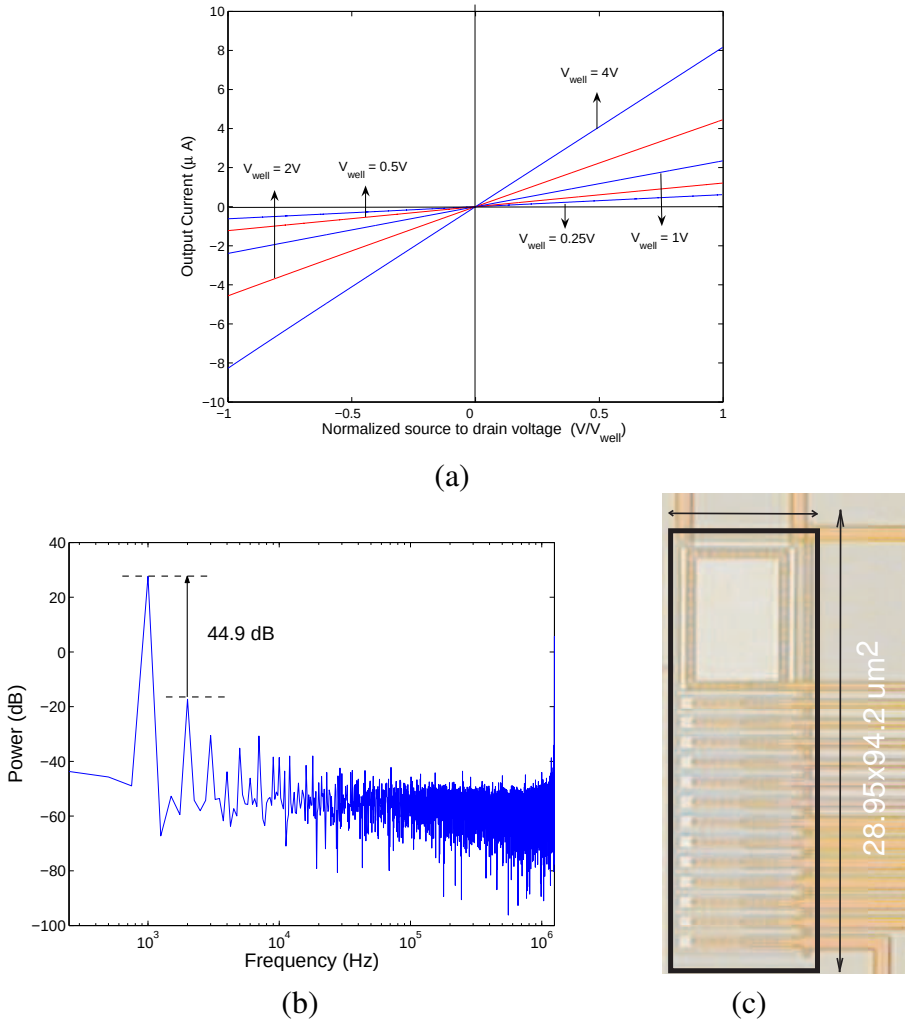
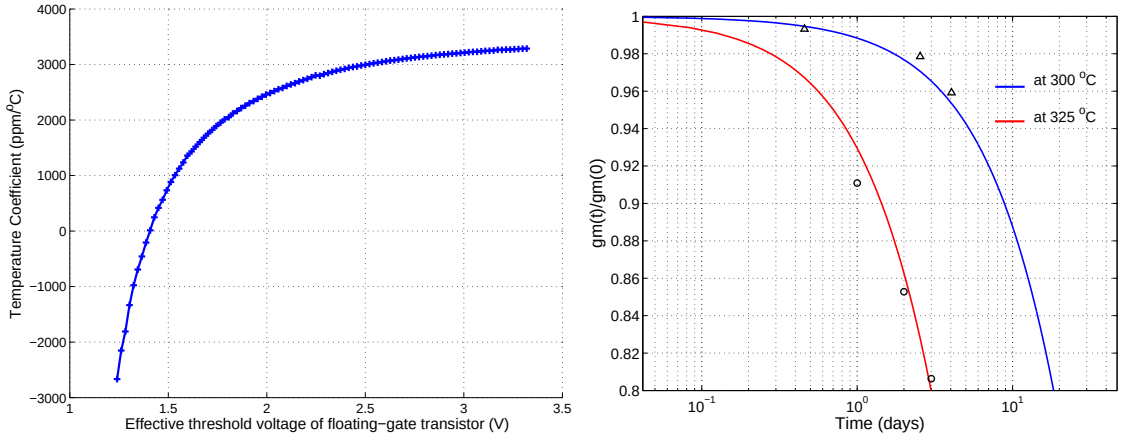


Figure 27. (a) Experimental results obtained with different well voltages, V_{well}. $x-axis$ of the plot is normalized to show the relative change. (b) The linearity test of the FGR_{SGL} for $1KHz$ sinusoidal input signal with $1V_{pp}$ amplitude. (c) Die photo of the FGR_{SGL}.

The dynamic measurements of the FGR_{SGL} are shown in Figure 27b, and obtained by using an off-chip inverting amplifier with corresponding feedback resistor. $1kHz$ sinusoidal wave with $1V_{pp}$ amplitude is used for the test, and the scale factor of the resistor is set to 0.7918. The second order harmonic distortion of the resistor for this test is measured as $44.9dB$. This is mainly because the second-order distortion of the FGR_{SGL} is not completely cancelled with the chosen scale factor.

Furthermore, the die photo of the fabricated FGR_{SGL} circuit is shown in Figure 27c. In the designed circuit, M_R has a dimension of $W/L = 19.5\mu m/1.2\mu m$. Also, the main



(a) (b)

Figure 28. (a) Temperature coefficient of the FGR_{SGL} for differently programmed threshold voltages. (b) Stress test of the FGR_{SGL} performed at 300^oC and 325^oC.

capacitor is $560fF$ and each trimming capacitor is $56fF$. The total area of the test circuit is $2727\mu m^2$.

The FGR_{SGL} temperature coefficient for a range of programmed effective threshold voltages is illustrated in Figure 28. The effective threshold voltages of the FGR_{SGL} are obtained from their gate sweeps. It is observed that this coefficient can be changed from $-2500ppm/^oC$ to $3300ppm/^oC$. Moreover, the long-term drift is mainly caused by the thermionic emission [62]. The resistance change over time can be found by using the following equation

$$\frac{g_m(t)}{g_m(t_0)} = \Phi(t,T) + \frac{\beta V_T}{g_m(t_0)}\Big[\Phi(t,T) - 1\Big] \tag{24}$$

where $\Phi(t,T) = exp[-tv.exp(\frac{-\phi_B}{kT})]$, $g_m = 1/R$ is the conductance, v is a relaxation frequency of electrons in poly-silicon, ϕ_B is the $Si - SiO_2$ barrier potential, k is the Boltzmann's constant. Figure 28 illustrates the stress test results. The worst case results are obtained after the first stress test at 300^oC. After the first test, the charge loss of the FGR_{SGL} is decreased considerably. The ϕ_B and v from these experiments are extracted as $0.9eV$ and $60s^{-1}$. Based on this worst-case data, it is calculated that the FGR_{SGL} resistance drifts $1.6 \cdot 10^{-3}\%$ over the period of 10 years at 25^oC.

To sum up, in this chapter, an implementation of a tunable floating-gate CMOS resistor

is presented. This resistor exploits the floating-gate transistor properties and the scaled-gate linearization technique. Better than $7-bit$ linearity is obtained for $1V_{pp}$ sinusoidal input. The circuit does not consume additional power for the offset and feedback generation, thus becomes very suitable for ANN systems and low-power applications. Furthermore, we showed that for a fixed scale factor, the well voltage can be reduced while still preserving the linearity of the resistor. Therefore, this CMOS resistor can be easily integrated with low-voltage applications.

CHAPTER 5

TUNABLE HIGHLY LINEAR FLOATING-GATE CMOS RESISTOR USING COMMON-MODE LINEARIZATION TECHNIQUE

The linearity and operating range of the resistors are the most crucial features for highly linear applications that require high signal-to-noise and distortion. In this work, we propose a tunable CMOS resistor that can be suitably employed in highly linear circuits. This CMOS resistor operates in the triode region, utilizes the common-mode linearization technique [18], and achieves a compact and power efficient circuit implementation by employing floating-gate MOS transistors. In the next section, we explain the common-mode linearization strategy, and analyze its effect. Subsequently, we describe the implementation of a tunable floating-gate resistor using the common-mode linearization (FGR_{CML}). After that, we present the experimental results of this circuit. In the last part of this chapter, we compare this resistor with previously reported resistors and discuss their characteristics.

5.1 Common-mode Linearization Technique

There are three principal nonlinearities in the drain current of a long-channel transistor in the triode region and these are identified as the body effect, the mobility degradation, and the fundamental quadric component due to the common-mode of the drain and source voltages. These nonlinearities are mostly dependent on the common-mode of the input signals, and can be suppressed by building common-mode feedback structures around a transistor [18].

The common-mode linearization scheme is illustrated in Figure 29, and exploits the fact that the linearity of a single transistor can be greatly improved by applying the common-mode signal (with the addition of their corresponding quiescent voltages) to the gate and body terminals [18]. Similar to the gate linearization, this technique also requires $v_{ds} < 2(V_G - v_s - V_T)$ to operate in the triode region, where v_{ds} is the drain-to-source voltage, V_G

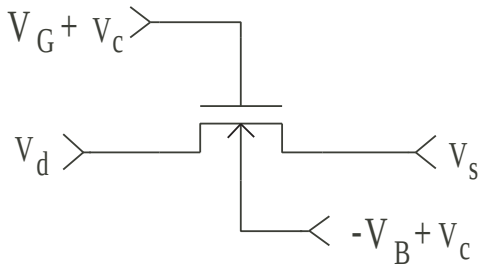


Figure 29. Common-mode linearization technique [18] applied to an nMOS transistor in the triode region. This method allows to minimize the nonlinearities of a MOS transistor by modulating the body and gate terminals with the common-mode voltage. v_d and v_s are the drain and source voltages, respectively. V_G and V_B are the tunable quiescent gate and body voltages, and v_c is the common-mode voltage, $v_c = (v_d + v_s)/2$.

is the quiescent gate voltage, v_s is the source voltage, and V_T is the threshold voltage. In this technique, the gate and body voltages, v_g and v_b, are defined as

$$v_g = V_G + v_c, \quad v_b = -V_B + v_c \tag{25}$$

where V_B is the quiescent body voltage, and v_c is the common-mode voltage and equal to $(v_d + v_s)/2$. Also, V_T, θ_2, and μ_2 are defined as

$$V_T = V_{FB} + \phi + \gamma\sqrt{V_B + \phi} \tag{26}$$

$$\theta_2 = \frac{\theta}{1 + \theta\Big(V_G - V_{FB} - \phi + \gamma\sqrt{V_B + \phi}\Big)} \tag{27}$$

$$\mu_2 = \frac{\mu_0}{1 + \theta\Big(V_G - V_{FB} - \phi + \gamma\sqrt{V_B + \phi}\Big)} \tag{28}$$

where V_{FB} is the flat-band voltage, ϕ is the surface potential, γ is the body-effect coefficient, θ is the mobility degradation factor, and μ_0 is the carrier mobility. As suggested in [18] and explained in Appendix-I, by using the above equations, the drain current for $\theta_2 \ll \frac{96(V_B+\phi)^{3/2}}{\gamma v_{ds}^3}$ can be approximated as

$$I_d = \frac{\mu_2 C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} + \frac{\gamma(1 + \theta_2[V_G - V_T])}{96\sqrt[3]{V_B + \phi}}v_{ds}^3\right\} \tag{29}$$

where C_{ox} is the gate capacitance per unit area, W is the channel width, and L is the channel length. The above result is remarkable in the sense that the inherent nonlinearities of a MOS transistor can be reduced down to a cubic ordered term. With a reasonable selection

of the quiescent gate and bulk voltages, the linear region of a MOS transistor can be greatly extended. After ignoring the higher order terms, the resistance of the linearized element can be expressed as

$$R = \frac{L}{\mu_2 C_{ox} W (V_G - V_T)} \tag{30}$$

In the above equation, V_T does not depend on the common-mode of the input voltages, thus the resistance exhibits suppressed common-mode voltage dependence.

5.2 Circuit Implementation

The circuit implementation of the common-mode linearization technique is shown in Figure 30. FGR_{CML} operates as a tunable floating resistor by exploiting the features of the floating-gate transistors. The common-mode voltage of the input signals is computed by using the feedback capacitors, which couple the drain and source voltages to the gate terminal. In addition, the charge stored on the floating-gate terminal creates the required quiescent gate voltage to satisfy the triode condition and linearity requirement. As shown in Figure 30a, V_{tun} is used to enable the tunnelling mechanism to decrease the number of electrons on the floating-gate terminal of M_R. Also, V_{sPROG} and V_{dPROG} are used to create the required voltage difference that is necessary for the hot-electron injection mechanism to occur and increase the number of electrons at the gate terminal of M_R. As a result, by using (2) the floating-gate voltage can be expressed as

$$V_{fg} = \frac{(C_g + C_{gs})V_s}{2C_g + C_{gs} + C_{gd} + C_{M_P} + C_{tun}} + \frac{(C_g + C_{gd})V_d}{2C_g + C_{gs} + C_{gd} + C_{M_P} + C_{tun}} + V_p \tag{31}$$

where V_p is the effect of the stored charge and the capacitive coupling from the peripheral circuit that includes C_{tun} and C_{M_P}. C_{tun} is the tunnelling junction capacitance, and C_{M_P} is the input capacitance of the injection transistor, M_P. C_{gs} becomes equal to C_{gd} for large quiescent gate voltages. Therefore, the necessary condition for an accurate common-mode computation is to create a large quiescent gate voltage and to keep C_g much larger than C_{M_P}

and C_{tun} so that the floating-gate potential is close to

$$V_{fg} \simeq \frac{(V_s + V_d)}{2} + V_p \tag{32}$$

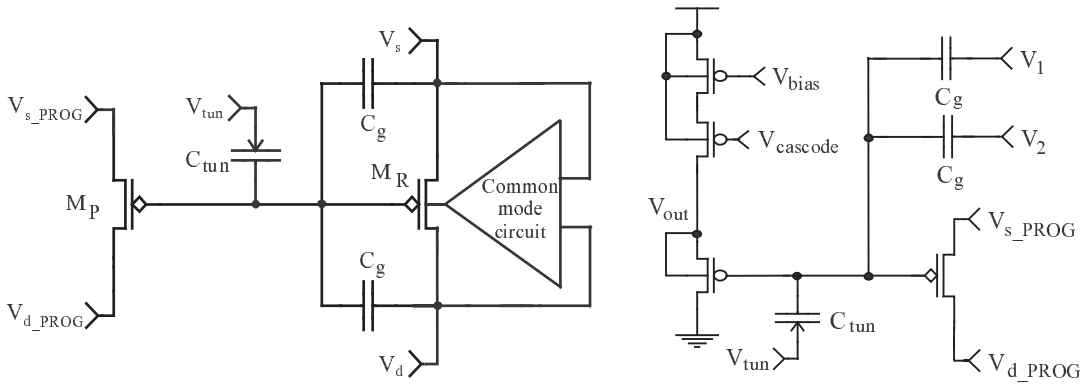
The scaling error introduced by the common-mode computation increases the common-mode dependence of the circuit. The circuit, shown in Figure 30a, employs a well feedback in addition to the gate feedback to further reduce the inherent nonlinearities of a MOS transistor. The common-mode computation circuit is illustrated in Figure 30b. This circuit is a source follower and used to drive the well terminal of M_R. Similar to the gate common-mode computation, two capacitors are used to compute the common-mode voltage at the input of the source follower. Since the well voltage has to be larger than the input voltages, V_d and V_s, to prevent the drain and source junctions from being forward biased, an offset voltage must be created at the input of the follower. This is achieved by programming (in this case by tunnelling) the charge stored on the floating gate of the follower. In addition, if a rail-to-rail operation is required with this resistor, the source follower needs to be powered with a higher supply voltage to accommodate the output voltage swing.

For an accurate common-mode well feedback voltage computation, the well feedback capacitors (C_g) have to be sized relative to the input capacitance of the source follower. In this case since the input transistor of the source follower operates in the saturation region, the input capacitance approximately becomes $C_{gs} = 2C_{ox}A/3$, where A is the total area of the input transistors.

If there is a mismatch between the gate feedback capacitors, or if there is a scaling error, then an error term, ε, is introduced to (29). This error can be approximated as $\varepsilon(v_d^2 - v_s^2)/2$, which is equal to $\varepsilon v_{ds} v_c$. Then, the drain current can be approximated as

$$I_d = \frac{\mu_2 C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} \pm \frac{\varepsilon}{2}(v_d + v_s)v_{ds} + \frac{\gamma(1 + \theta_2[V_G - V_T])}{96\sqrt[3]{V_B + \phi}} v_{ds}^3\right\} \tag{33}$$

As a result, the error term in the drain current gives rise to a common-mode voltage dependence. In modern processes, a matching accuracy better than 0.1% can be obtained with the



(a) (b)

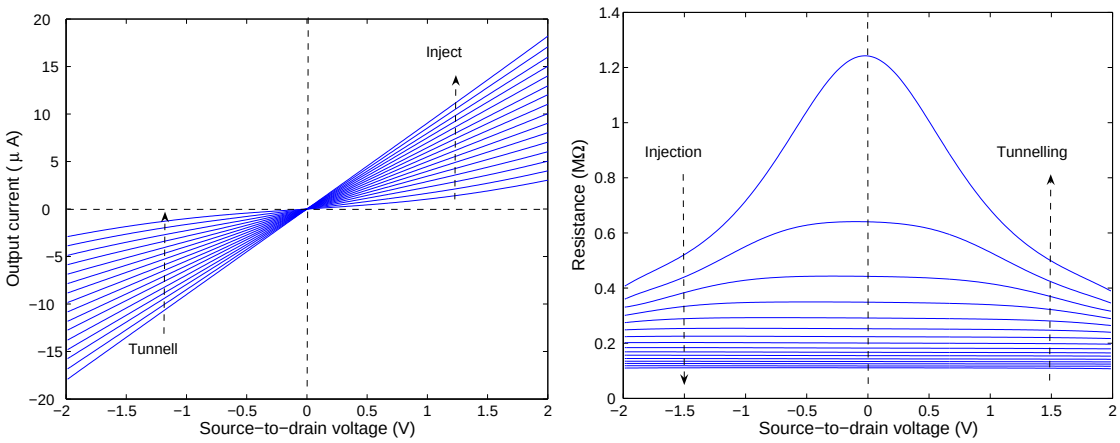
Figure 30. (a) Circuit implementation of the tunable floating-gate resistor. V_s and V_d are the source and drain voltages of M_R, respectively. This resistor is tuned by changing the quiescent gate voltage. This is achieved by using the tunnelling junction connected to V_{tun}, and the injection transistor that has source voltage V_{sPROG} and drain voltages V_{dPROG}. The feedback capacitors (C_g) are used to compute the common-mode gate voltage. Also, the well feedback voltage is computed by the common-mode circuit. (b) The common-mode computation circuit. This circuit consists of a source follower, a programming circuitry and input capacitors (C_g). Input capacitors compute the common-mode voltage and apply it to the input of the buffer. V_{bias} is used to set the current through the circuit, and $V_{cascode}$ is employed to minimize the effect of the output voltage on the bias current. The computed common-mode voltage is tracked by the buffer circuit and then applied back to the well.

capacitors, and this together with high quiescent gate voltages readily allow for the circuit implementation of the CMOS resistors with high linearity.

5.3 Experimental results

In this section, we present the characterization results of the proposed circuits. The measurements are obtained from the chips that were fabricated in a 0.5μm CMOS process. A 16-bit DAC is used for the measurements to characterize the linearity and voltage-dependence of the FGR_{CML}.

The experiments for the static measurements are performed by keeping one terminal of the floating-gate resistors at 2.5V, and then sweeping the other terminal between 0.5 and 4.5V. Also, the source follower of the FGR_{CML} is powered with 6V during the experiments. After the each programming step by tuning the quiescent gate voltage, the experiment is repeated to observe the change in the resistance and linearity. This is achieved by tuning

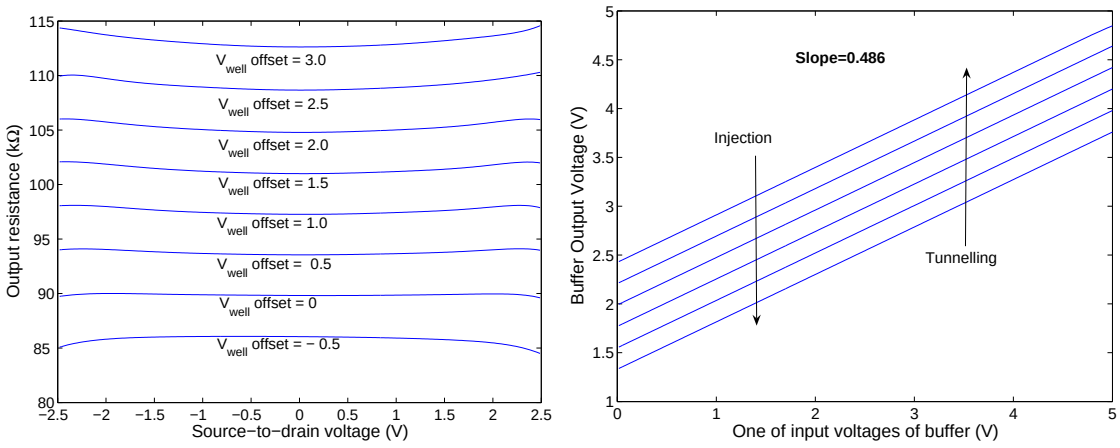


(a) (b)

Figure 31. Experimental results. The measurements are performed by keeping one of the terminals at $2.5V$ and sweeping the other terminal from $0V$ to $5V$. These measurements are obtained for differently tuned quiescent gate voltages, which is increased through injection to decrease the resistance. Also, this gate voltage is decreased by using tunnelling to increase the resistance. (a) The output current vs. input voltage sweeps. (b) The resistance vs. input voltage sweeps. Extracted resistances of the FGR_{CML} tuned to different quiescent gate voltages.

the amount of stored charge on the floating-gate terminal of M_R. The I-V curves of the FGR_{CML} are shown in Figure 31a. The FGR_{CML} exhibits less variation for its smaller resistance values as shown in Figure 31b. This is mainly because the relative effect of the common-mode voltage on the resistance becomes less for higher V_G values, and V_T stays almost fixed in the operating range of the resistor. In addition, having a larger V_G helps the transistor to stay in the deep triode region even for large differential input signals. Moreover, the nonlinearities of this structure are a function of the quiescent gate voltage, and they can be better suppressed for large gate quiescent voltages. This is especially true for the FGR_{CML}, since θ_2 becomes very small for large gate quiescent voltages.

The well voltage of M_R affects the resistance and the linearity of the FGR_{CML}. When the offset voltage of the source follower is increased from $-0.5V$ to $3V$, the resistance of the FGR_{CML} changes $\pm 15\%$ as shown in Figure 32. Here, V_{well} offset is defined as the voltage difference between the source/drain and well voltages of M_R when the drain/source voltage is $5V$. For that purpose, the source follower is powered with $9V$ to observe the resistance variation of the FGR_{CML} when one of its input terminals swept from $0V$ to $5V$ and the

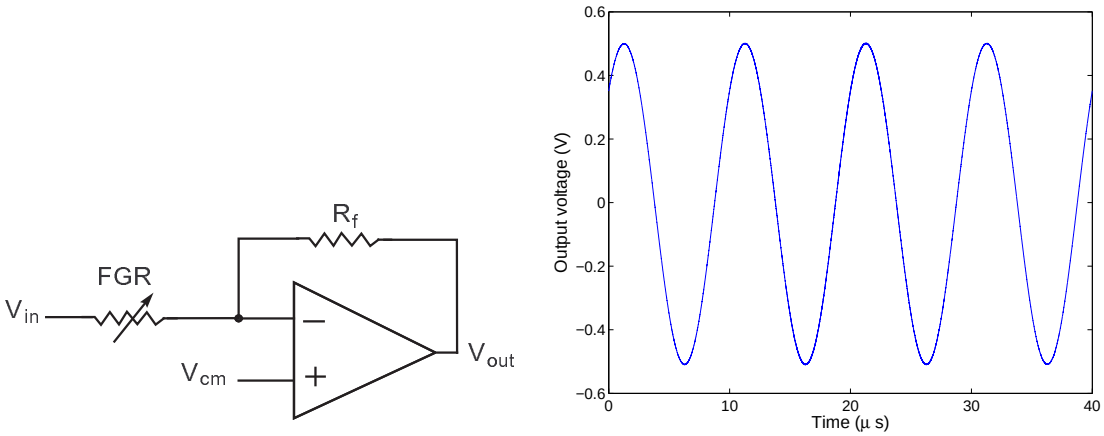


(a) (b)

Figure 32. (a) Effect of the well offset voltage on the resistance of the FGR_{CML}. (b) Output voltage of the source follower when one of the FGR_{CML} inputs is swept from $0V$ to $5V$. The slope is measured to be 0.486.

other terminal is fixed at $2.5V$. The output voltage of the source follower and its offset programming using the Fowler-Nordheim tunnelling and hot electron injection is depicted in Figure 32b. It is observed that the slope of the well common-mode computation is only 0.486 and not 0.5. This difference causes asymmetry in the output current of the FGR_{CML} and increases the second-order harmonic distortion of the FGR_{CML}.

The dynamic measurements of the FGR_{CML} are obtained by using an off-chip inverting amplifier with a corresponding feedback resistor (matches the resistance of on-chip resistor) as shown in Figure 33a. For this purpose, a sine-wave with $2.5V$ offset and $1V_{pp}$ amplitude is used to test the transient behavior as well as the distortion level of the FGR_{CML}. The maximum frequency of the input signal that can be used with the FGR_{CML} depends on the resistance and the input capacitance of the FGR_{CML}. It is also important to use enough bias current for the source follower so that it can drive the well terminal of M_R at given frequency. The FGR_{CML} transient response for $100kHz$ and $1V_{pp}$ input sine-wave is shown in Figure 33b. For this transient test, the source follower is biased with $10\mu A$. Moreover, the total nonlinearity and harmonic distortion of the FGR_{CML} are tested using this test setup and utilizing the 16-bit DAC. It is observed that the total nonlinearity of the FGR_{CML} with $W/L = 1.2/7.5$ in the full operating range of $\pm 2V$ can be held below 1% while changing its



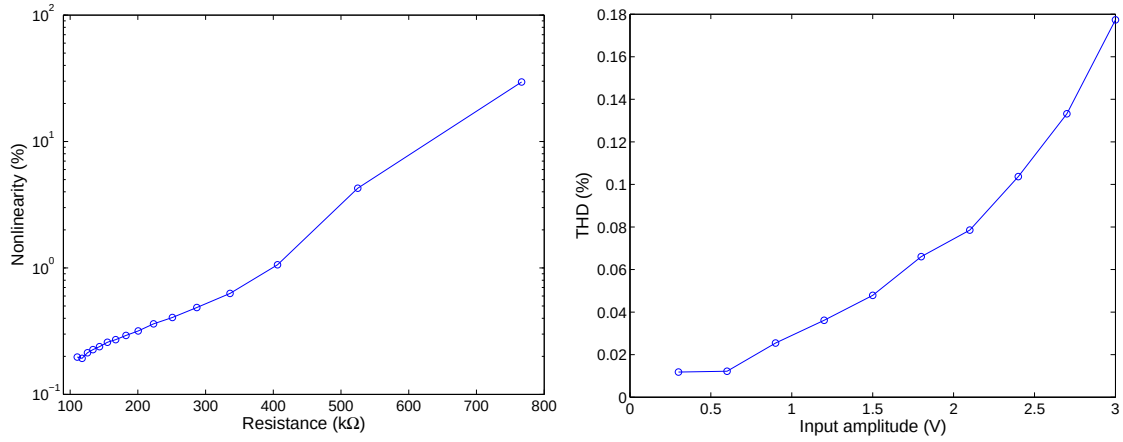
(a) (b)

Figure 33. (a) Inverting amplifier used to test the transient behavior and distortion level of the FGR_{CML}. (b) Transient measurement data of the FGR_{CML} for $100kHz$ and $1V_{pp}$ sine-wave.

resistance from $100k\Omega$ to $600k\Omega$ as shown in Figure 34a.

When the FGR_{CML} is tuned to have a resistance around $100k\Omega$, its total harmonic distortion for sine amplitude levels from $0.3V$ to $3V$ increases from 0.012% to 0.18% as illustrated in Figure 34b. Although, a better linearity performance is possible with this structure, due to inaccurate computation in the well computation circuit the distortion level of the FGR_{CML} is measured to be higher. Since the common-mode computation by the source follower results in 0.486, especially the second-order harmonic of the FGR_{CML} output increases considerably. This reasoning is justified by testing the FGR_{CML} nonlinearity and its THD for a range of well feedback ratios. This test is performed by supplying the well feedback potential of M_R from off-chip 16-bit DAC.

As shown Figure 35a, the nonlinearity of the FGR_{CML} in the full operating range can be reduced below 0.1% when the well feedback ratio is very close to 0.5. Similarly, the THD of the FGR_{CML} for $1V_{pp}$ input signals becomes around 0.005% for well feedback ratio of 0.5. Therefore, the well feedback ratio is found to be the main source of error in this resistor structure. Furthermore, the well offset voltage also affects the linearity of the FGR_{CML}. The second and third-order harmonic distortions of the FGR_{CML} is also a function of the well offset voltage as illustrated in Figure 36a. The third-order harmonic

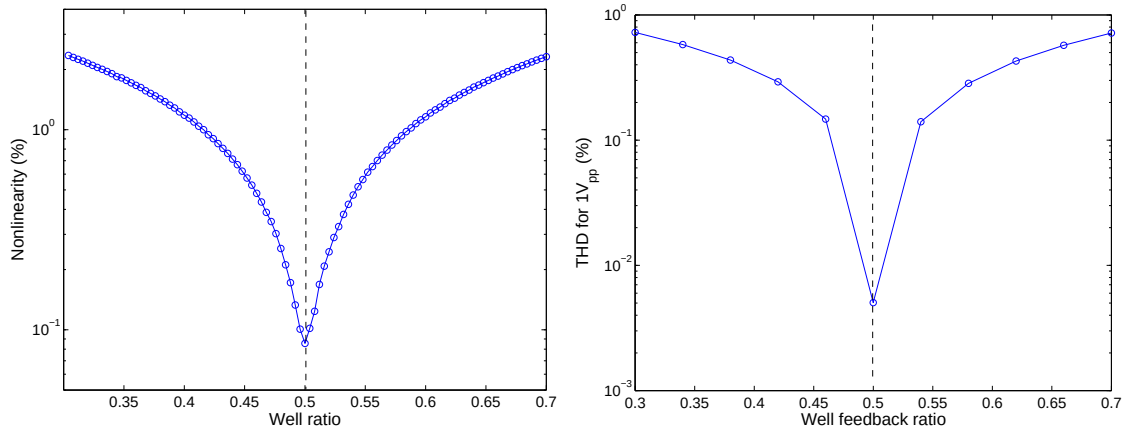


(a) (b)

Figure 34. a) Nonlinearity of the FGR_{CML} for differently tuned resistance values. The nonlinearity is measured in the full operating range of $\pm 2V$. The resistance is tuned by changing the quiescent gate voltage, which is increased through injection (to decrease the resistance) and decreased by using tunnelling (to increase the resistance). (b) Total harmonic distortion of the FGR_{CML} for a sine-wave with different amplitude levels.

distortion can be reduced from $-60dB$ to $-95dB$ when the well offset is increased from $0.75V$ to $2.25V$. However, the second-order harmonic distortion is mainly caused by the inaccuracy of the well feedback ratio, thus it does not change much with the well offset voltage.

Lastly, the change of the FGR_{CML} resistance with the input voltage is tested for smaller transistor lengths. Since the initial assumption in building this structure is to have a long-channel transistor, the linearity of the FGR_{CML} decreases and its resistance changes much more for smaller channel lengths as depicted in Figure 36b. The die photo of the fabricated chip is shown in Figure 37. The dimensions of M_R is $W/L = 1.5\mu m/15\mu m$, and the values of each gate and well feedback capacitors are $450fF$ and $1970fF$, respectively. These capacitors can be optimized depending on the input capacitance of the transistors. Also, an auxiliary bias generator circuit is used to generate the bias current and the cascode voltage to be used for the source follower.



(a) (b)

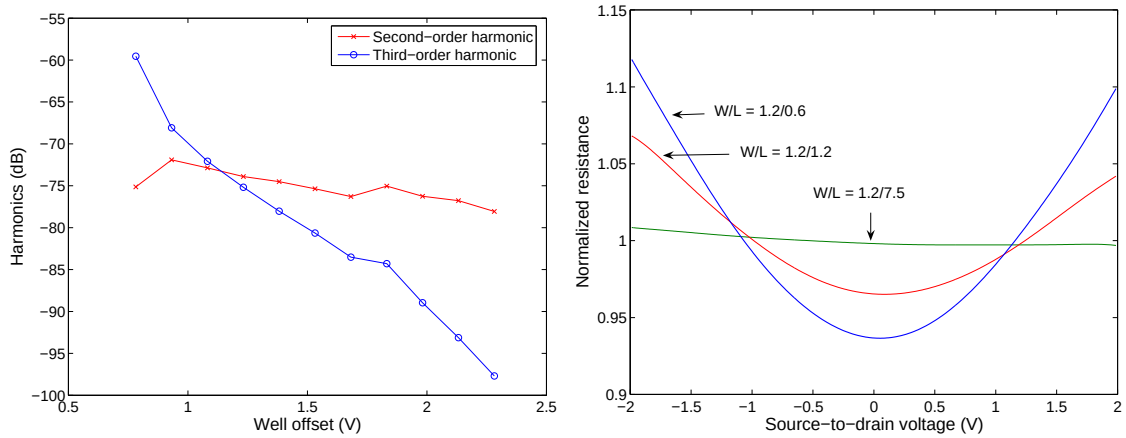
Figure 35. Measurement results for a range of well feedback ratios. The well potential of M_R is supplied from off-chip. (a) The nonlinearity of the FGR_{CML} in the full operating range of $\pm 2V$ for a range of well feedback ratios. (b) Total harmonic distortion of the FGR_{CML} for $1V_{pp}$ input sine wave for a range of well feedback ratios.

5.4 Discussion

The results obtained from the presented CMOS resistor make this structure very suitable for variety of applications. In Table 1, the characteristics of other resistor implementations are summarized to compare FGR_{CML} with these implementations. These resistors are implemented in BICMOS or CMOS processes.

A resistor implementation exploiting the square-law characteristics of the transistors has a resistance that is independent of the threshold voltage of the CMOS transistors [7]. This floating resistor achieves 1% THD for $2.4V_{pp}$. It is implemented with 20 transistors, and it allows tuning from $56k\Omega$ to $112k\Omega$ (can be scaled for chosen W/L). The main shortcoming of this implementation is that it requires the use of large area due to the number of transistors employed, but does not offer high linearity.

Moreover, a voltage controlled MOS resistor based on the bias-offset technique operates within the 80% of the supply range, and achieves $\pm 1\%$ THD for $8V_{pp}$ input signals [12]. 9 transistors are used to implement this compact MOS resistor. Although this resistor offers a compact implementation, it is prone to second order effects caused by the channel-length modulation, mobility degradation, and device mismatches.



(a) (b)

Figure 36. (a) The second and third-order harmonics of the FGR_{CML} for a range of well offset voltages. The well offset voltage is changed by programming the offset voltage of the source follower by using the injection and tunnelling mechanisms. (b) Normalized resistance of the FGR_{CML} circuits vs. their input voltage. The length of M_R is sized as $0.6\mu m$, $1.2\mu m$, and $7.5\mu m$.

In addition to these implementations, a CMOS resistor structure based on the current division technique [68] allows for high linearity even with large voltage swings. It yields 0.01% THD for $2.5V_{pp}$ signals. However, 4 transistors, 1 amplifier, and 4 resistors increase the area overhead and the power consumption of this implementation.

Furthermore, a 6-terminal CMOS resistor [69] implemented in a BICMOS process achieves the best linearity performance within the reported resistors. While around 0.0032% THD is possible with this structure for $1V_{pp}$ input signals, its size and power consumption are the main disadvantages of this design. Also, the BICMOS process increases the cost of this implementation compared to its counterparts in CMOS processes.

The floating-gate resistor that is reported in this work utilize the properties of MOS transistors in a CMOS process. FGR_{CML} has 4 transistors, 4 capacitors in addition to the two programming circuitry. This structure results in increased linearity, which is necessary for the most of highly linear applications and reduced power consumption since only one source follower needs to be powered. At most $72dB$ (for $1V_{pp}$) of linearity is obtained with this FGR_{CML} design. It is observed that the accuracy of the well feedback computation is the limiting factor for the resistor linearity. Also, the implementation of the FGR_{CML} in

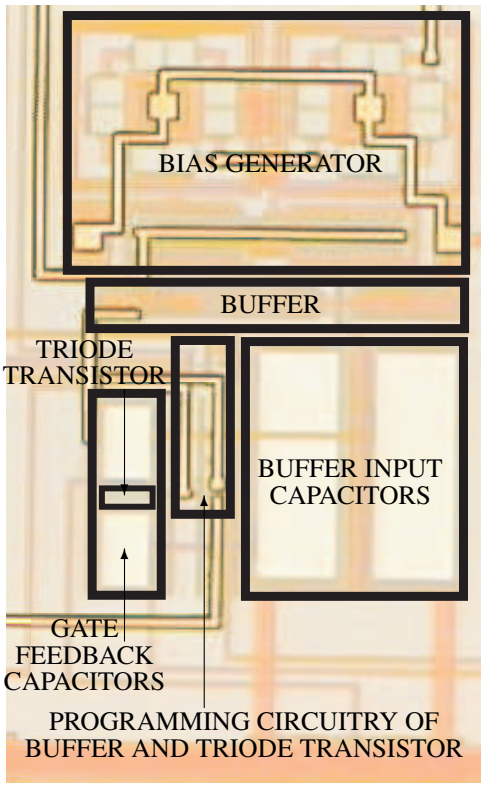


Figure 37. Die photo of the fabricated FGR_{CML} circuit.

CMOS processes with feature sizes smaller than $0.35\mu m$ can be achieved by using thick-oxide transistors if available. In this chapter, we presented an implementation of tunable CMOS resistor by making use of the floating-gate transistor features. We showed that the tuning and operating ranges of the resistor are extended by employing the analog storage characteristic of the floating-gate transistors. Also, we showed that FGR_{CML} offers a compact and power efficient implementation that yields around $72dB$ of linearity. The linearity and power efficiency of this resistor make it suitable for highly linear circuit applications.

Table 1. Experimental results of tunable CMOS resistors (T:transistor, R:resistor, C:capacitor, B:buffer, LS:level shifter, A:amplifier, PC:programming circuitry)

Design	[7]	[12]	[68]	[69]	FGR_{CML}
Process	2μm CMOS	3μm CMOS	2μm CMOS	BICMOS	0.5μm CMOS
Power supply	10V	10V	5V	-	6V
Operating range	2.4V	8V	3V	10V	4V
Tuning range	56 to 112 $k\Omega$	-	$\pm 5\%$	-	100 to 800 $k\Omega$
THD	1% (2.4V_{pp})	$\pm 1\%$ (8V_{pp})	0.01% (2.5V_{pp})	$< 0.0032\%$ (1V_{pp})	0.024% (1V_{pp})
Components	20T	9T	4T+1A+4R	1T+4B+4LS	4T+4C+2PC

CHAPTER 6

DESIGN OF HIGHLY LINEAR AMPLIFIER AND MULTIPLIER CIRCUITS USING A CMOS FLOATING-GATE RESISTOR

The linearity of the highly linear amplifier and multiplier circuits can be increased by employing the highly linear tunable CMOS resistor described in Chapter 5. This resistor can serve as an alternative to passive resistors and allow the realization of a dynamic and linear resistor while facilitating a reduction in system size and cost. In the next section, we explain how this resistor can be used to increase the linear range in differential amplifiers and to implement two-quadrant transconductance multipliers. In the last part of this chapter, we present the experimental results of these circuits.

6.1 Highly Linear Amplifier Design

The highly linear amplifier circuit is shown in Figure 38a. This circuit is implemented by using high gain amplifiers to achieve the voltage-to-current conversion without introducing additional distortion. Also, it employs the FGR_{CML} circuit as a variable resistor, R_{var}. Each high gain amplifier consists of an input differential amplifier and folded-cascode output stage that results in a high gain [70]. With the use of these amplifiers, NMOS current mirrors achieve boosted g_m and conduct the current (I_{p3}+I_{p2}/2) plus the signal current i_s, which is the current created by the differential voltage, v_{in} across R_{var}. When the finite open-loop gain, A_0, of the amplifiers is taken into account, the signal current can be expressed as

$$i_s = \frac{v_{in}}{R_{var} + 2/(g_m(1 + A_0))} . \frac{A_0}{1 + A_0} \tag{34}$$

This equation shows that for more accurate voltage-to-current conversion and less distortion, high gain is required. In order to prevent the capacitive loading at the resistor stage and to improve the frequency response and linearity of the circuit, the feedback capacitors

of the FGR_{CML} are buffered by employing the same source follower used for the well feedback shown in Figure 38b. Similar to the gate common-mode computation of the FGR_{CML}, two gate capacitors are used to compute the common-mode voltage at the input terminal of the source follower. This structure employs a highly linear source follower [71] to drive the well terminal. This open-loop source follower is preferred because of its wider bandwidth than the closed loop followers and its high linearity. Since the well voltage has to be larger than the input voltages, V_d and V_s, to prevent the drain and source junctions from being forward biased, an offset voltage must be created at the input of the follower. This is achieved by programming (in this case by tunnelling) the charge at the follower gate input terminal enough to obtain the voltage needed for the operation of the resistor. Additionally, if a rail-to-rail operation is required, then a higher supply voltage needs to be used for the source follower.

For this application, the folded cascode amplifier is preferred over grounded amplifiers [72] not only to obtain a higher gain but also to avoid the additional V_{sg} drop that can counteract the effect of the injected charge at the gate of the pMOS floating-gate resistor. In addition, the input transistors of the amplifier are chosen to be pMOS to utilize their n-well for eliminating the body effect and improving the noise performance.

6.2 Multiplier Design

MOS transistors in the triode region can be used to implement transconductance multipliers. A two-quadrant multiplier circuit is designed by using a single FGR_{CML} circuit as shown in Figure 38c. One of the source/drain terminals of the floating-gate transistor is fixed and used as an output, V_{out}, and the other terminal is employed as an input, V_d. The second input of the multiplier, V_r, is supplied from the feedback gate capacitor, C_{g1}. While a bidirectional current is created by utilizing V_d, this current is modulated by changing the conductance of the FGR_{CML}. For this purpose, the FGR_{CML} has to be put into the triode regime by injecting enough electrons to the gate terminal so that the transistor stays in the

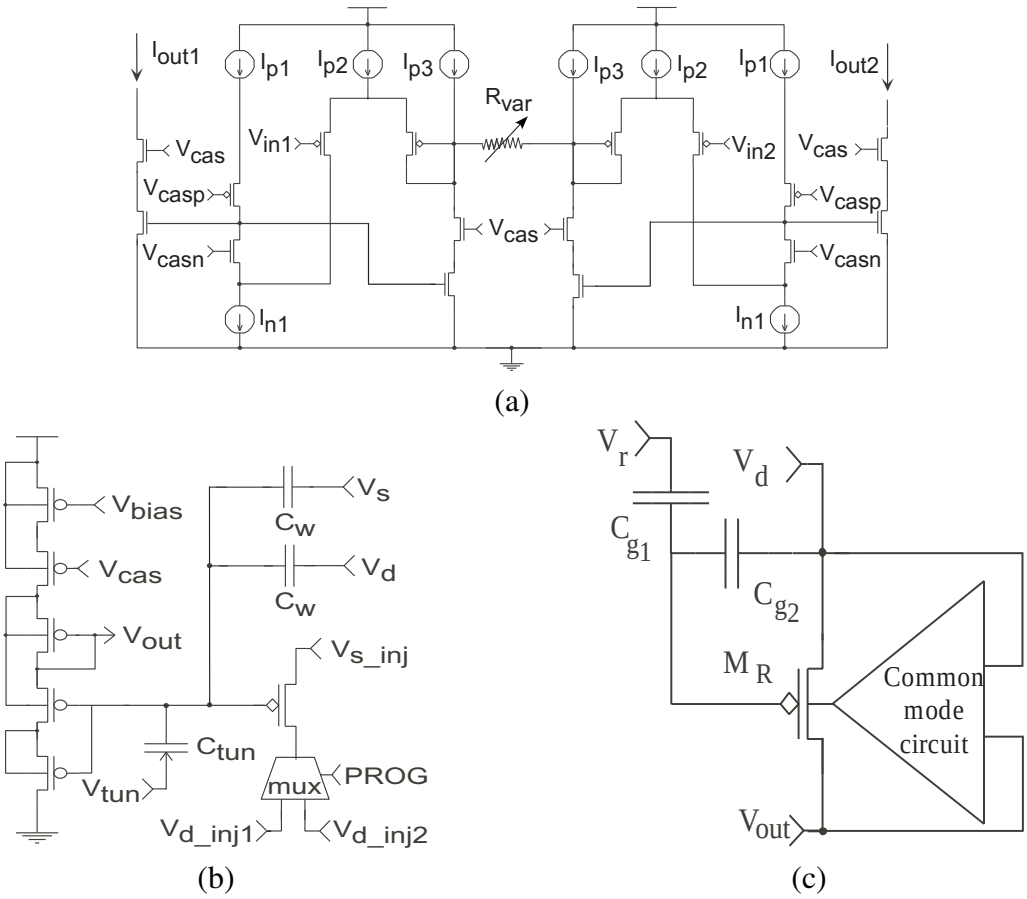
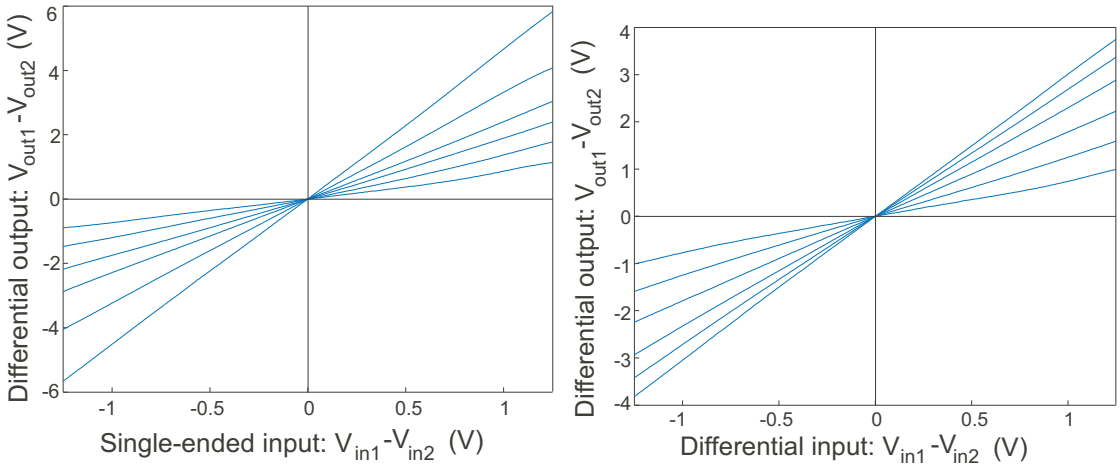


Figure 38. (a) Circuit implementation of the variable gain amplifier. The FGR_{CML} is used as a variable resistor, R_{var}. (b) The common-mode computation circuit. It consists of a highly linear source follower [71], programming circuitry and input capacitors. Input capacitors compute the common-mode voltage and apply it to the input of the follower. The computed common-mode voltage is tracked by the follower circuit and applied to the well. (c) Two quadrant multiplier circuit implementation. V_r and V_d are the input voltages and V_{out} is the output voltage, and the output of the circuit is obtained in the form of current.

linear region for the required input swing. In addition, the gate capacitor modulates the resistance of the circuit by changing the effective voltage at the gate terminal, and this can be shown by ignoring the higher-order terms in the FGR_{CML} current,

$$I_{out} = \frac{\mu_o' C_{ox} W}{L}[\frac{V_r}{2} + V_G - V_T](V_d - V_{out}) \tag{35}$$

where V_G in this equation is defined as the effect of the charge at the gate and capacitor couplings when V_r is set to V_{out}. Assuming the common mode voltage of V_r is V_{out}, then the amplitude of V_r needs to be smaller than $2(V_G - V_T)$ so that multiplier stays in the triode region.



(a) (b)

Figure 39. Experimental results of the highly linear amplifier. The output current of the amplifier is converted to voltage by using $10K\Omega$ on-chip resistors. Input-output DC characteristics of the amplifier for differently tuned FGR_{CML} values. (a) Differential output response of the amplifier to a single-ended input. V_{in1} is used as a input while V_{in2} is kept constant at $2.5V$. (b) Differential output response of the amplifier to a differential input.

This multiplication gives two terms, $V_r(V_d - V_{out})$ and $(V_G - V_T)(V_d - V_{out})$. The second term can be removed by using two multiplier circuits, and then by applying fully differential signals to their input capacitors. In this case, multipliers must be programmed to the same resistance value for accurate offset cancellation. The subtraction of output currents of the multipliers results in a four-quadrant multiplication, and the output can be expressed in terms of $(V_{r_1} - V_{r_2})(V_d - V_{out})$, where V_{r_1} and V_{r_2} are the differential inputs.

6.3 Experimental results

In this section, we present the characterization results of the proposed circuits. The measurements are obtained from the chips that were fabricated in a 0.5μm CMOS process.

The DC characteristics of the highly linear amplifier for single-ended and differential inputs are shown in Figure 39a and 39b , respectively. It is shown that it is possible to apply $2.5V_{pp}$ single-ended and differential inputs. This range is mainly limited by the cascode transistors of the amplifier. The output current of the highly linear amplifier is converted to

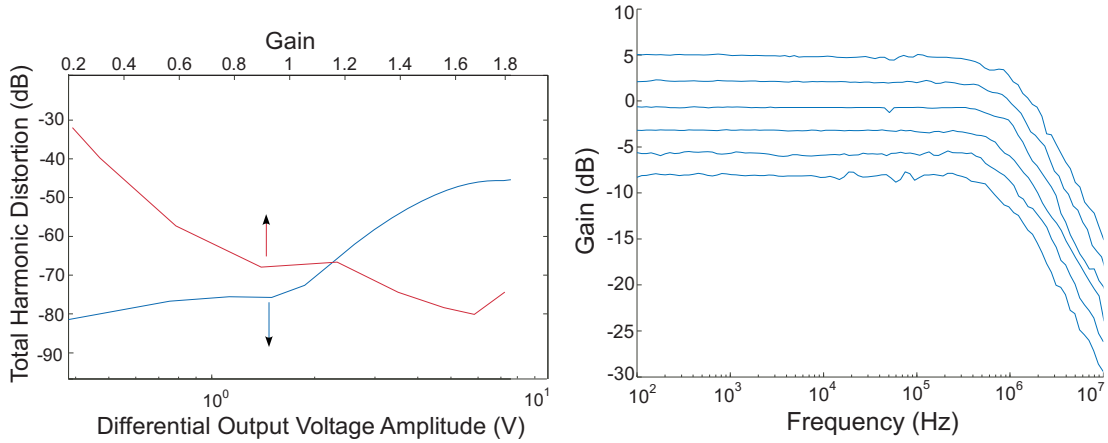


Figure 40. Experimental results of the highly linear amplifier. (a) Total harmonic distortion of the amplifier for differential input signals. The upper curve represents the total harmonic distortion of the amplifier for different gains, which is defined as $10K\Omega/R_{FGR}$ in this context. Output voltage amplitude is fixed at $1V_{pp}$ for distortion measurements and the gain is changed by tuning the FGR_{CML} to different resistance values. The lower curve illustrates the distortion levels of the amplifier for a range of output voltage amplitudes. For this measurement, the gain is fixed at 1.5 by tuning the FGR_{CML}. (b) The frequency response of the amplifier for different gains obtained by tuning the resistance of the FGR_{CML}.

voltage by using on-chip $10K\Omega$ resistors, and then buffered for off-chip reading. Total harmonic distortion (THD) of this amplifier for a range of signal amplitude and amplifier-gain is illustrated in Figure 40a. The amplifier can yield 0.018% THD for $1V_{pp}$ differential input. Increase in the input voltage amplitude and in the FGR_{CML} resistance cause degradation in the linearity of the amplifier.

Furthermore, the frequency sweeps of the amplifier for differently tuned FGR_{CML} resistance values are shown in Figure 40b. The amplifier has a $3dB$ frequency around $1MHz$, and this limitation is mainly caused by the buffer circuit as well as breadboard parasitics. The performance of the amplifier is summarized in Table II.

Finally, the dynamic results of the multiplier circuit is illustrated in Figure 41. The output current of the multiplier is converted to voltage for off-chip reading. It is shown that the output of the multiplier fits well with the theoretical results. The linearity and linear range of the multiplier can be improved by increasing V_G in (35) since FGR_{CML} in the multiplier circuit becomes more linear. There are two design issues with the FGR_{CML}

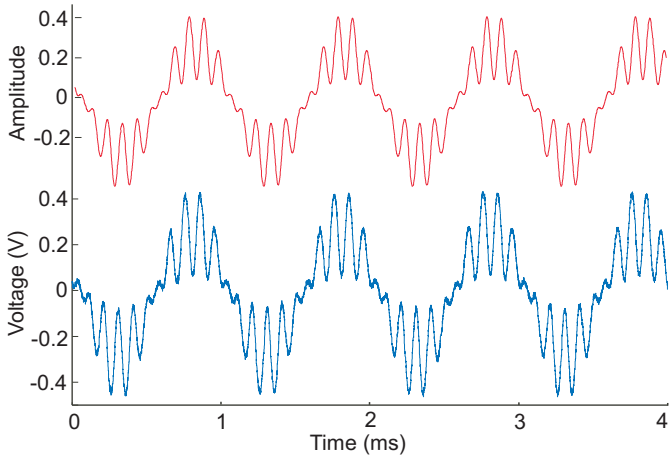


Figure 41. Output of the multiplier to a $1KHz$, $1V_{pp}$ input signal while its gate is modulated with $10KHz$, $1.5V_{pp}$ signal. The upper curve is a theoretical result of the multiplication and the lower curve illustrates the output of the multiplier. Theoretical result shows that the response of the multiplier fits with the equation $sin(w_0t + \phi_0) \cdot (2.475 + 1.5 \cdot sin(10w_0t))$, where ϕ_0 is the phase difference between two input signals.

structure. Firstly, the source follower has to operate with larger power supply voltages than V_{dd} if a rail-to-rail resistor operation is required. Secondly, the feedback capacitors has to be large enough to minimize the effect of the peripheral circuit. The parasitic capacitors and finite matching of the feedback capacitors may prevent the accuracy in the common-mode voltage computation.

In this chapter, it is shown that a tunable resistor can be employed to design highly linear amplifier and two-quadrant multiplier circuits. Also, the design of a four-quadrant multiplier circuit is described. The amplifier exhibited 0.018% THD for $1V_{pp}$ differential input, and a linear input range of $2.5V_{pp}$. These circuits will be employed in applications where the linearity and tuning ranges are primary concerns.

Table 2. Experimental Performance of the Amplifier

Power supply	5V
Power consumption	5mW
ICMR	≃2.5V
Gain range for $THD > -55dB$ (single FGR_{CML})	-5dB to 5dB
$3dB$ frequency	≃1MHz
Technology	0.5μm CMOS
Active die area	0.06 mm^2

CHAPTER 7

DESIGN OF A BINARY-WEIGHTED RESISTOR DAC USING TUNABLE LINEARIZED FLOATING-GATE CMOS RESISTORS

In this chapter, the design of a binary-weighted resistor DAC using the linearized tunable resistor (FGR_{SGL}) is described. This tunable resistor is implemented in a standard CMOS process and provides a high resolution and precise device calibration through the use of floating-gate transistors. In contrast to previously reported floating-gate CMOS resistors [67] [66], this resistor has a simple structure and provides a high degree of design flexibility in optimizing the overall area and the tuning range of the DAC. In the next section, we describe the design and implementation of the binary-weighted resistor DAC. Subsequently, we present the experimental results of this circuit.

7.1 Design and implementation of binary-weighted resistor DAC

The binary-weighted resistor DAC structure is depicted in Figure 42. Variable resistors are used to obtain the scaled currents and full output voltage swing at the DAC output. The input resistors, R_i for $i = 1, ..., N$, switch between ground and voltage reference, V_{ref}, and generate the scaled currents. Also, V_c and R_c are used to obtain a larger output voltage range by creating an offset current. In addition, due to tunability of these resistors, R_c enables to tune the offset of the DAC.

In this kind of implementation, where accuracy is the main design objective, highly matched passive resistors are used in the design to prevent any degradation in the DAC linearity. However, this requirement necessitates the use of large devices, which can be expensive in terms of area and may degrade the high frequency performance. Instead of passive devices, tunable resistors can be used to alleviate matching and area requirements.

Ideally, this structure is immune to resistor non-linearity since, to a first approximation, the voltage across the resistors can assume only two values. However, due to the limited

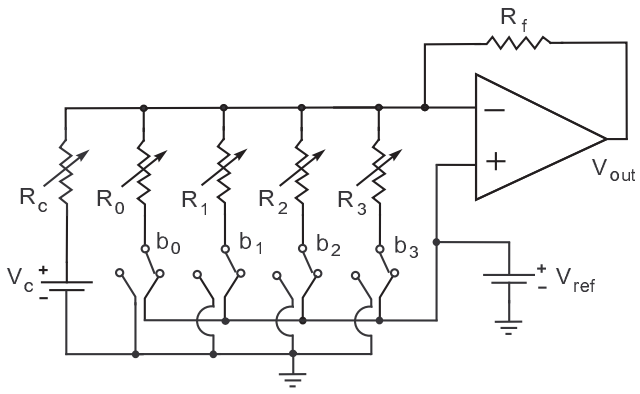


Figure 42. Proposed implementation of a binary-weighted DAC using tunable resistors. R_i is the tunable resistor, where $i = 0, 1, 2, 3$. Also, R_f is the feedback resistor, and R_c is used to obtain the full output voltage range and to tune the offset of the DAC. V_c is set to supply rail of the DAC.

low frequency voltage gain of the amplifier, the voltage across the resistors still vary by the error voltage, $e = V_o/A_o$, where V_o is the output voltage swing and A_o is the low frequency voltage gain of the amplifier. Therefore, when tunable resistors are incorporated into such design, the nonlinearities of these resistors have to be suppressed to obtain a better DAC linearity.

In a standard CMOS technology, a tunable CMOS resistor can be designed by using an MOS transistor operating in the triode region. However, MOS transistors operating in the triode-mode exhibit a large resistance variation mainly due to their quadratic dependence on voltage across their source and drain terminals. For this reason, it is necessary to apply a linearization technique to MOS transistors to enable their use as a variable resistor in DACs. Therefore, FGR_{SGL} shown in Figure 24 is utilized in this DAC implementation to obtain the scale factors. Use of floating-gate transistors in this DAC structure enables to obtain the tunable scale factors.

Due to the asymmetric structure of the FGR_{SGL}, one of its input terminals has to be maintained at a fixed potential. Hence, V_s terminals of these resistors are connected to the corresponding switches while their V_d terminals are connected to the inverting node of the amplifier. In this resistor structure, V_{g_2} can be used to tune the resistance of the FGR_{SGL}.

As long as the FGR_{SGL} stays in the triode region, V_{g_2} can alter the transconductance of the FGR_{SGL} linearly since it has a linear relation with the effective gate voltage.

7.2 Experimental results

In this section, we present the measurement results of the proposed circuits that were fabricated in a 0.5μm CMOS process. The input capacitors, C_{g_1} and C_{g_2}, are sized as $2016fF$ and $784fF$, respectively, to obtain a scale factor of $\chi = 0.72$. The scaled resistors are implemented by using scaled transistors with $W = 1.2\mu m$, and $L = 2.4\mu m$, $4.8\mu m$, $9.6\mu m$, and $19.2\mu m$.

The DC characteristics of the FGR_{SGL} circuits are obtained by keeping their drain terminal at ground and sweeping their source terminals from 0 to $5V$ as illustrated in Figure 43a. In this experiment, the well potential is fixed to $5V$. The extracted resistances of these resistors are shown in Figure 43b, where resistances are scaled by the scale factor of the resistors to observe the relative change in their resistance. The precise scale factors for the implementation of the DAC are obtained by tuning the resistance of the FGR_{SGL} for a source-to-drain voltage of $2.5V$, which is the reference voltage of the DAC. As the length of the tunable resistor increases, the deviation in the FGR_{SGL} resistance decreases. This is mainly because the scaled-gate linearization technique becomes more effective for the long channel devices. The temperature dependence of the FGR_{SGL} is shown in Figure 43c and obtained by changing the temperature from -40 to 80^oC. The temperature coefficient of the FGR_{SGL} is measured as $2770ppm/^oC$.

The static characteristics of the 4-bit DAC are illustrated in Figures 44. DAC has an output voltage range of $4.56V$, and the INL and DNL plots illustrate that the accuracy error can be limited to less $139\mu V$, which corresponds to 15-bit of accuracy. The MSB step response of the DAC is shown in Figure 45a, and depending on the size of the feedback capacitor, settling time less than $10\mu s$ can be obtained. The sine wave test is shown in Figure 45b. $1kHz$ sinusoidal signal is generated by setting the sampling frequency at $170kHz$.

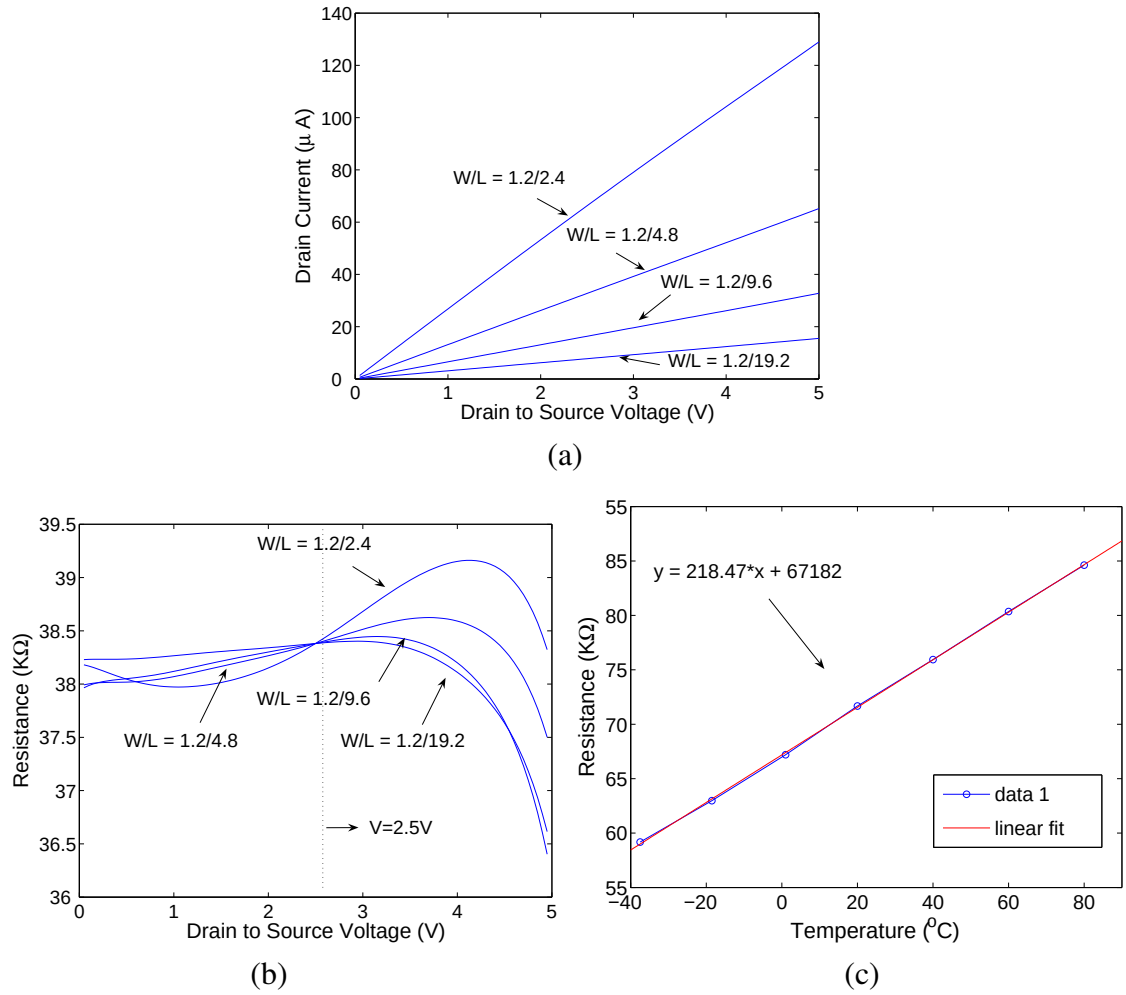


Figure 43. (a) Voltage sweeps of the tunable resistors from 0 to $5V$. (b) Extracted resistances of the tunable resistors with different lengths. For visual purposes, all other resistances are scaled to $W/L = 1.2\mu m/2.4\mu m$. (c) Temperature sweep of the FGR_{SGL} for $W/L = 1.2\mu m/4.8\mu m$.

This DAC can be made much faster by properly sizing the FGR_{SGL}.

The long-term and short-term drift of the DAC is crucial as it determines the DAC reliability. The short-term drift can be observed shortly after the floating-gate programming, and can be minimized by decreasing the number of injection pulses for the fine tuning of the devices. The short-term drift of the DAC linearity is illustrated in Figure 45c. It is observed that after programming the DAC for 15-bit accuracy, the linearity drops to around 14-bit. Moreover, the long-term drift of the DAC resistors is mainly caused by the thermionic emission [62]. Based on the stress tests, it is calculated that the FGR_{SGL} resistance drifts

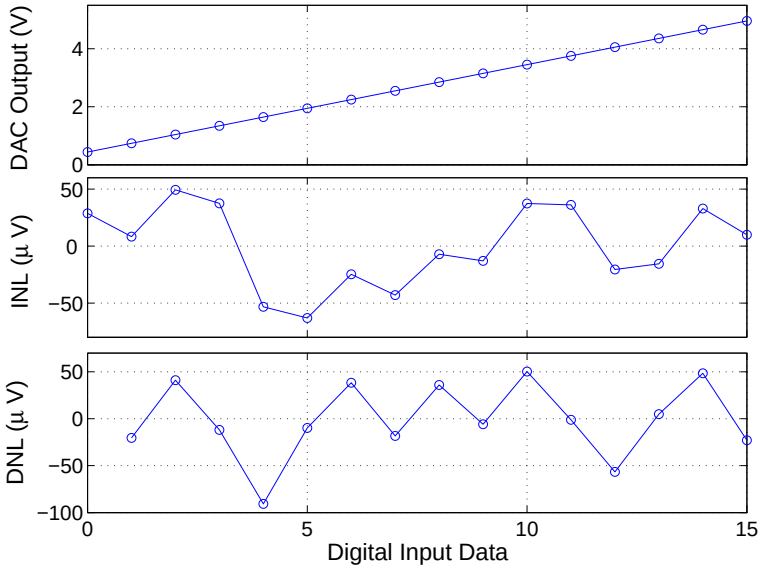


Figure 44. Static characteristics of the DAC: Output voltage, INL, and DNL.

$1.6 \cdot 10^{-3}\%$ over the period of 10 years at $25^{o}C$.

In this chapter, the implementation of a binary-weighted resistor DAC using tunable floating-gate CMOS resistors is presented. It is shown that the resistance and temperature coefficient of the FGR_{SGL} can be tuned to a desired operating point. The stress test of these resistors showed that the FGR_{SGL} resistance drifts negligibly over time. It was also demonstrated that $15-bit$ accurate, $4-bit$ resolution DAC can be built using these resistors. This will readily enable the implementation of multi-bit CMOS quantizers in pipelined and over-sampling data converters.

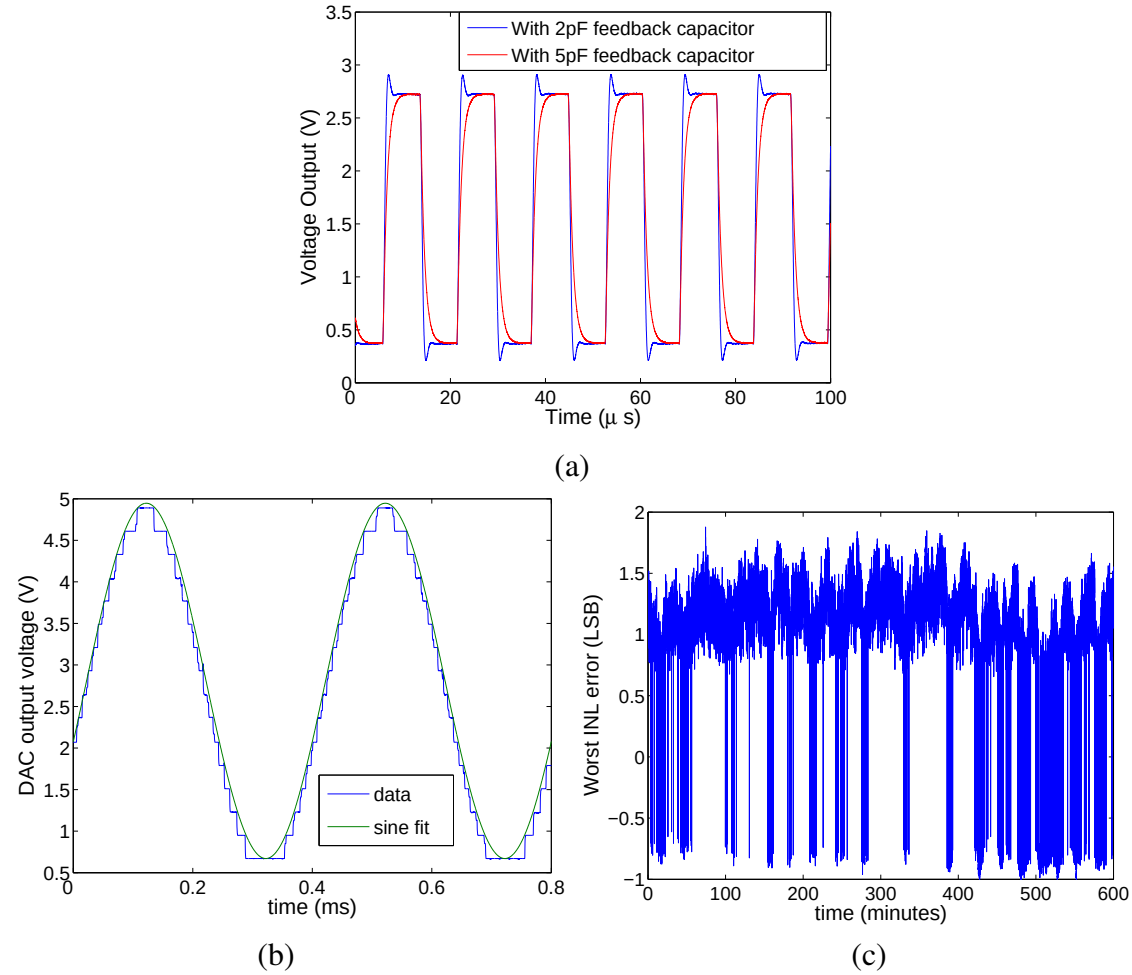


Figure 45. (a) MSB step responses for $2pF$ and $5pF$ feedback capacitors. (b) Sinusoidal transient response of the DAC. The sinusoidal-fit is shown to illustrate the behavior of the DAC response. (c) Short term linearity test of the DAC. The 10-hour data illustrates the change of the linearity over time for $LSB = 139\mu V$.

CHAPTER 8

PROGRAMMABLE VOLTAGE-OUTPUT DIGITAL-TO-ANALOG CONVERTER

Floating-gate transistors can be utilized to obtain a better performance optimization for Nyquist rate converters that require low-power and small area. In this chapter, we propose the use of programmable floating-gate voltage references (epots) to build a floating-gate based binary-weighted DAC (FGDAC). The epot is an ideal device for obtaining a dynamically reprogrammable, non-volatile, on-chip voltage reference in standard CMOS processes [73]. Utilizing epots to compensate for capacitor mismatches and to obtain binary-weighted voltage levels enable to implement a DAC with an unity element spread. This implementation results in a compact, low-power voltage-output DAC. Earlier results [74, 75] demonstrated the feasibility of the epot integration into a charge amplifier architecture.

In the next section, the binary-weighted capacitor DAC (BWCDAC) is compared with the FGDAC, and their area, speed, accuracy, and noise performances are compared. Subsequently, the circuit architecture of the FGDAC is explained, and integration of epots into this implementation is described. In the last part of this chapter, the experimental results of the FGDAC are presented.

8.1 Traditional binary-weighted capacitor vs. proposed DAC design: BWCDAC vs. FGDAC

In the traditional design of the BWCDAC, capacitors are scaled, and additional switches are incorporated to periodically clear the inverting node of the amplifier as illustrated in Figure 4. This structure has its own limitations mainly due to its scaled capacitor array. Some of the trade-offs and limitations of the BWCDAC can be alleviated by utilizing the FGDAC implementation, shown in Figure 46. This implementation employs epots to obtain the scaled voltage levels, which readily allow for a fixed-area-per-bit. In addition, the reset switches in the traditional design can be eliminated by using floating-gate transistors to

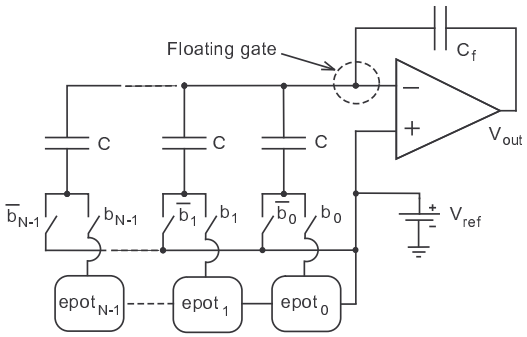


Figure 46. Proposed design of floating-gate based DAC (FGDAC) that uses scaled voltages instead of scaled capacitors to achieve the digital-to-analog conversion. In this design, C_f is equal to C. This converter is implemented by employing epots in a charge amplifier structure. Reference voltages for each bit are programmed both to scale the input voltages and to minimize the effect of the mismatch between capacitors.

control the charge on the inverting node of the amplifier. Therefore, return-to-zero phase in the BWCDAC design can also be eliminated and the timing requirements of the DAC can be relaxed. Other than these differences, the analysis for the DAC area, speed, gain error, and noise performances are provided in the following subsections to show the design improvements and trade-offs.

8.1.1 Area

The area allocated for the capacitor array of the BWCDAC depends on the unity-size-capacitor area, A_C, and on the number of bits, N. Therefore, the total capacitor area used for this converter becomes $A_{C_f} + A_{C_s} = (2^{N+1} - 1) \cdot A_C$, where A_{C_f} and A_{C_s} are the area used for feedback capacitor and scaled capacitors, respectively. In this equation, $C_f = 2^N C$ and $C_s = (2^N - 1)C$.

In contrast, the total area used for obtaining the scale factors of the FGDAC is mostly determined by the epots. Therefore, the total area increases linearly with the number bits, and can be computed as $A_{C_f} + A_t = (N + 1) \cdot A_C + N \cdot A_{epot}$, where A_t is the total area used for the input capacitors and epots, and A_{epot} is the area of an epot.

As a result, to obtain an improvement in the total DAC area using the FGDAC implementation, A_{epot} has to be smaller than $A_C \cdot (2^{N+1} - N - 2)/N$ for an N-bit converter. For

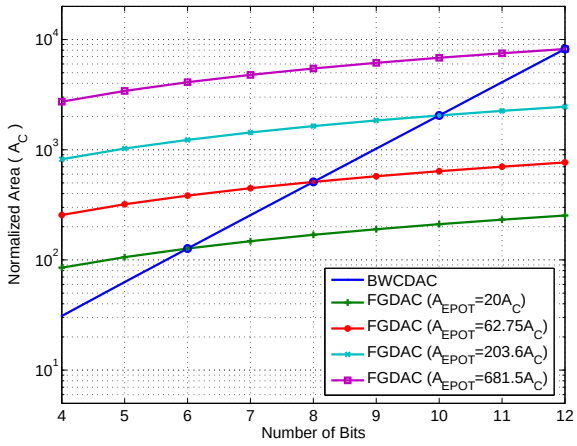


Figure 47. Comparison of the BWCDAC and the FGDAC for the area used to achieve the binary-weighted scaling. This area corresponds to the area of the capacitor array for the BWCDAC, while it is the sum of the areas of the capacitor and epot arrays for the FGDAC. A_C **and** A_{EPOT} **are the capacitor and epot areas. The FGDAC area is computed for a range of epot area,** $A_{EPOT} = \alpha \cdot A_C$**, where** $\alpha = 20, 62.75, 203.6, 681.5$**.**

high resolution converters, the FGDAC implementation becomes an advantageous design approach to minimize the total DAC area. The areas of the BWCDAC and the FGDAC are compared for a range of A_{epot} values, as shown in Figure 47. The curves in this plot represent only the total area used for scaling, and exclude the area used for other DAC components. The intersection of these curves represent the point where the areas of the BWCDAC and the FGDAC become equal for given number of bits and epot area. Therefore, the FGDAC design strategy can yield a more compact converter depending on the value of A_{epot} and the number of bits. For instance, the total capacitor area of the $10-bit$ DAC can be reduced around 100 times for $A_{epot} = 20 \cdot A_C$ if same size unit capacitors are used in building the BWCDAC and the FGDAC.

8.1.2 Speed

The speed of the BWCDAC and the FGDAC are compared based on their time constants. Here, it is assumed that these converters are structurally same. Since the time constants of these converters are dependent on the type of the amplifier, one and two-stage amplifier models are used to compare the converter speeds.

The time constants of the BWCDAC and the FGDAC are defined as τ_{BWCDAC} and

τ_{FGDAC}, respectively. For unit capacitance, C, the feedback and the total input capacitance of the BWCDAC are $C_f = 2^N C$ and $C_{eq} = (2^N - 1)C$. However, these capacitance values become $C_f = C$ and $C_{eq} = NC$ for the FGDAC. Moreover, in this analysis, the output resistance of the voltage references are assumed to be very small compared to the on-resistance of the switches.

8.1.2.1 Using one-stage amplifier

Based on the analysis given in the Appendix-II, the time constants, τ_{DAC_1} and τ_{DAC_2}, can be computed as

$$\tau_{DAC_1} = R_{on}C + \frac{C_f C_L + (C_{amp} + C_{eq})(C_f + C_L)}{G_m C_f} \tag{36}$$

$$\tau_{DAC_2} = \frac{R_{on}C(C_f(C_{amp} + C_L) + C_L C_{amp})}{C_f(G_m R_{on}C + C_L) + (C_{amp} + C_{eq})(C_L + C_f)} \tag{37}$$

where R_{on} is on-resistance of the switches, G_m is the amplifier transconductance, C_{amp} is the amplifier input capacitance, C_L is the load capacitance, and C_{eq} is the sum of the input capacitors.

When designing the converters, it is important to keep R_{on} small enough to utilize the full bandwidth of the amplifier. It can be shown that if $R_{on}C \ll (C_f C_L + (C_{eq} + C_{amp})(C_f + C_L))/(G_m C_f)$, then $\tau_{FGDAC_2} < \tau_{FGDAC_1}$ and $\tau_{BWCDAC_2} < \tau_{BWCDAC_1}$. Therefore, the first time constants of these converters determine their maximum speed. In this case, the ratio of τ_{BWCDAC_1} and τ_{FGDAC_1} can be expressed as

$$\frac{\tau_{BWCDAC_1}}{\tau_{FGDAC_1}} = \frac{C_L + (1 + C_{amp}/(2^N C))(C_L + 2^N C)}{C_L + (N + C_{amp}/C)(C_L + C)} \tag{38}$$

The relationship between the BWCDAC and the FGDAC speeds based on the above equation is illustrated in Figure 48. This equation indicates that for negligibly small amplifier input capacitance and for a small load capacitor the FGDAC operates much faster than the BWCDAC does.

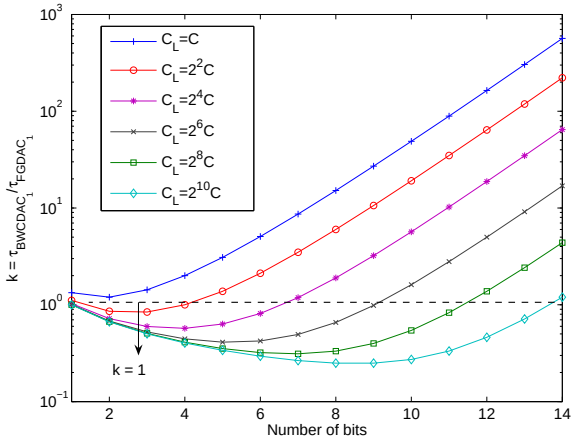


Figure 48. Speed comparison of the BWCDAC and the FGDAC for one-stage amplifier case and small amplifier input capacitance. The ratio of their time-constants, k, shows the relation for increasing number of converter bits. $k = 1$ represents the same speed performance for these converters. k is computed for $C_L = 2^{\lambda} \cdot C$, where $\lambda = 0, 2, 4, 8, 10$, to show the effect of the load capacitance on the BWCDAC and the FGDAC speeds. The FGDAC is faster than the BWCDAC for $k > 1$.

Table-3 summarizes all the cases based on the initial assumption that $R_{on}C$ is very small. According to these results, it can be concluded that when the FGDAC is used with one-stage amplifier, it performs better than the BWCDAC for $C \gg C_L$ and $C \gg C_{amp}$. The first condition necessitates the use of a buffer if the DAC is designed for off-chip purposes.

8.1.2.2 Using two-stage amplifier

The time constants of the BWCDAC and the FGDAC for a two-stage amplifier are computed by using the analysis in Appendix-II. Based on this analysis, the time constants, τ_{DAC_1} and τ_{DAC_2}, can be written as

$$\tau_{DAC_1} = \frac{1}{GB} \cdot \left(1 + R_{on}C \cdot GB + \frac{C_{eq} + C_{amp}}{C_f}\right) \tag{39}$$

Table 3. Speed comparison for one-stage amplifier case.

Capacitors	$\tau_{BWCDAC}/\tau_{FGDAC}$
$C \gg C_L$ & $NC \gg C_{amp}$	$2^N/N$
$C_L \gg C_{amp}$ & $C_{amp} \gg 2^N C$	$1/2^N$
$C_L \gg 2^N C$ & $NC \gg C_{amp}$	$2/(N+1)$
$C_{amp} \gg C_L$ & $C_L \gg 2^N C$	$1/2^N$
$C_{amp} \gg 2^N C$ & $C \gg C_L$	1

$$\tau_{DAC_2} = \frac{R_{on}C}{1 + \frac{C_{eq}+R_{on}C \cdot C_f GB}{C_f + C_{amp}}} \tag{40}$$

If $R_{on}C \gg \frac{1}{GB} \cdot (1 + \frac{C_{eq}}{C_f})$, the converter speeds become approximately equal. However, converters are designed not to be limited by the on-resistance of the switches. For this reason, it can be assumed that $R_{on}C \ll \frac{1}{GB} \cdot (1 + \frac{C_{eq}}{C_f})$ to help the speed comparison. As a result, the speeds of these converters are mostly determined by their first time constants. In this case, τ_{BWCDAC_1} and τ_{FGDAC_1} can be approximated as

$$\tau_{BWCDAC_1} \approx \frac{2 + C_{amp}/(2^N C)}{GB} \tag{41}$$

$$\tau_{FGDAC_1} \approx \frac{N + 1 + C_{amp}/C}{GB} \tag{42}$$

The ratio of τ_{BWCDAC_1} and τ_{FGDAC_1} becomes

$$\frac{\tau_{BWCDAC_1}}{\tau_{FGDAC_1}} = \frac{2 + C_{amp}/(2^N C)}{N + 1 + C_{amp}/C} \tag{43}$$

which implies that the BWCDAC is faster than the FGDAC by the factor determined by the number of bits. As the number of bits increases the BWCDAC performs better than the FGDAC in terms of speed. This is mainly caused by the fact that the feedback capacitor of the BWCDAC is much bigger than the feedback capacitor of the FGDAC, and this enables a better feedback factor for the BWCDAC. While C_f/C_{eq} is approximately one for the BWCDAC, it is $1/N$ for the FGDAC. However, it has to be noted that $\tau_{BWCDAC_1}/\tau_{FGDAC_1}$ becomes approximately one for $C_{amp} \gg 2^N C$.

8.1.3 Gain error

Due to finite gain, A_v, of the DAC amplifier, the BWCDAC has a gain error that can be computed using

$$\left|\frac{V_{out}}{V_{in}}\right| = \frac{C_{eq}}{C_f + \frac{C_f + C_{eq} + C_{amp}}{A_v}} \approx \frac{C_{eq}}{C_f}\left(1 - \frac{C_f + C_{eq} + C_{amp}}{A_v C_f}\right) \tag{44}$$

where the gain error is represented by the term, $(C_f + C_{eq} + C_{amp})/(A_v C_f)$. The gain error in the above equation increases as C_{eq}/C_f gets larger.

In contrast to the BWCDAC, the FGDAC does not suffer from the gain error as long as the gain stays constant in the bandwidth of interest. This is mainly because the voltage levels and the least-significant-bit (LSB) of the FGDAC can be set by using the stored epot voltages for a given amplifier gain.

8.1.4 Noise

In this section, the noise analysis of the BWCDAC and the FGDAC are presented for the DAC design with one-stage and two-stage amplifiers. Also, the individual noise contribution from the switches, the amplifier, and the references are compared for different capacitance values to investigate the optimum design approach for the FGDAC that can yield improved noise performance. In the bandwidth of interest, the total DAC noise can be written as

$$e_{DAC}^2 = e_{reset}^2 + e_{amp}^2 B_{n_1} A_{n_1} + (e_{ref}^2 + e_{Ron}^2) B_{n_2} A_{n_2} \tag{45}$$

where e_{amp}^2, e_{Ron}^2, and e_{ref}^2 are the broadband noise contribution of the amplifier, the switches, and the reference. Also, B_{n_1} and B_{n_2} are the noise bandwidths of the amplifier and the reference/switches, and A_{n_1} and A_{n_2} are the gain of the DAC from the amplifier and the reference/switches, respectively. e_{reset}^2 is the kT/C noise introduced during the reset phase of the BWCDAC. This reset noise does not exist in the FGDAC, since the FGDAC operates without resetting the inverting node of the amplifier.

The output noise of the BWCDAC can be computed using the noise contributions of the reset and amplification phases. During the reset phase, the feedback path of the amplifier is shorted, and all the capacitors are connected to the ground. The noise coming from the on-resistance of the switches during the reset phase is stored and added to the noise in the amplification phase. Therefore, by using the analysis in Appendix-III and assuming that N is large and $G_m R_{on} C \ll C$, the total thermal noise of the BWCDAC for one-stage amplifier can be approximated as

$$e_{BWC}^2 = \frac{kT}{2^N C} + \left(kT \frac{R_{on}}{2^N} + \frac{e_{ref}^2}{4} + e_{amp}^2 \right) \cdot \frac{G_m}{C_x} \tag{46}$$

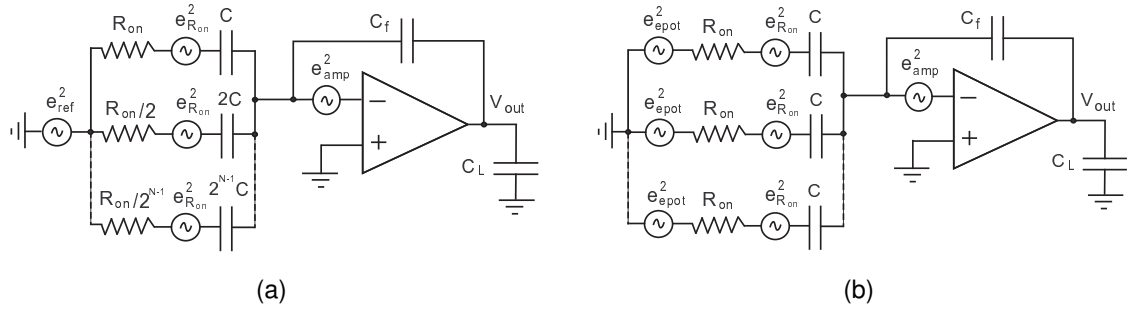


Figure 49. Simplified noise models of the BWCDAC and the FGDAC. $e^2_{R_{on}}$ and e^2_{amp} are the broadband contribution of the switches and amplifier. (a) Noise model of the BWCDAC during the amplification phase. For the worst-case analysis, all the capacitors are assumed to be connected to the reference voltage. e^2_{ref} is the noise contributions of the reference voltage. (b) Noise model of the FGDAC. e^2_{epot} is the noise contribution of the selected epot. Similar to the worst case analysis of the BWCDAC, all input capacitors of the FGDAC are assumed to be connected to their corresponding epots. Noise contribution of the reference voltage is ignored since it sets the common-mode of the amplifier and epots.

where $C_x = C_L + (1 + C_{amp}/(2^N C)) \cdot (2^N C + C_L)$. Similarly, for large values of N, the total thermal noise of the BWCDAC for two-stage amplifier becomes

$$e^2_{BWC} = \frac{kT}{2^N C} + (4kT\frac{R_{on}}{2^N} + e^2_{ref}) \cdot \frac{2^{N-2}C \cdot GB}{(2^{N+1}C + C_{amp})} + e^2_{amp} \cdot \frac{2^N C \cdot GB}{(2^N C + C_{amp})} \tag{47}$$

In contrast to the BWCDAC, the FGDAC has only one phase, where the selected voltage levels are summed for the digital-to-analog conversion. During the conversion, the total FGDAC noise is mainly contributed by the epots, the switches, and the amplifier. Based on the analysis in Appendix-III, and assuming that all of the epots are selected, N is large, and $G_m R_{on} C \ll C$, then the equivalent output thermal noise of the FGDAC for one-stage amplifier can be approximated as

$$e^2_{FG} = \left(4kT\frac{R_{on}}{N} + \frac{e^2_{epot}}{N} + e^2_{amp}\right) \cdot \frac{N^2 \cdot G_m}{4C_y} \tag{48}$$

where e^2_{epot} is the broadband noise contribution of the epot and $C_y = C_L + (N + C_{amp}/C)(C_L + C)$. For two stage amplifier, the equivalent thermal noise of the FGDAC becomes

$$e^2_{FG} = (4kTR_{on} + e^2_{epot}) \cdot \frac{NC \cdot GB}{4((N+1)C + C_{amp})} + e^2_{amp} \cdot \frac{(N+1)^2 C \cdot GB}{4(C + C_{amp})} \tag{49}$$

The total noise of the BWCDAC and the FGDAC for one-stage and two-stage amplifiers can be compared based on the individual contributions from the thermal noise of

the switches, the reference/epots, and the amplifier when the on-resistance, the amplifier transconductance, the load capacitance, the input capacitance, and the unit capacitance of these converters are the same.

To begin with, for one-stage amplifier, if $C_x \gg G_m R_{on} C$, the ratio of noise contribution due to the switches of the BWCDAC and the FGDAC can be expressed as

$$a_1 = \frac{1}{G_m R_{on}} \cdot \frac{C_L C + (NC + C_{amp})(C_L + C)}{2^N N C^2} \tag{50}$$

For two-stage amplifier and $GB \ll 1/(R_{on}C)$, this ratio becomes

$$a_2 = \frac{1}{GB \cdot R_{on} C} \cdot \frac{(N+1)C + C_{amp}}{2^N NC} \tag{51}$$

Similar to noise ratio of switches, the ratio of noise contributions from the amplifier for one-stage amplifier can be expressed as

$$b_1 = \frac{2^{N+2}}{N^2} \cdot \frac{C_L C + (NC + C_{amp})(C_L + C)}{2^N C_L C + (2^N C + C_{amp})(2^N C + C_L)} \tag{52}$$

and this ratio for two-stage amplifier can be written as

$$b_2 = \frac{2^{N+2}}{(N+1)^2} \cdot \frac{C + C_{amp}}{2^N C + C_{amp}} \tag{53}$$

Moreover, the ratio of noise contributions from the reference and epots for one-stage amplifier becomes as follows

$$c_1 = \frac{2^N}{N} \cdot \frac{C_L C + (NC + C_{amp})(C_L + C)}{2^N C_L C + (2^N C + C_{amp})(2^N C + C_L)} \tag{54}$$

For two-stage amplifier this ratio can be approximated as

$$c_2 = \frac{2^N}{N} \cdot \frac{(N+1)C + C_{amp}}{2^{N+1} C + C_{amp}} \tag{55}$$

To sum up, the above equations show that the total FGDAC noise due to the on-resistance of switches, the amplifier, and the references is comparable to the total noise of the BWCDAC. The BWCDAC exhibits better noise performance in some cases mainly

due to scaling difference between the feedback and input capacitors of the BWCDAC and FGDAC. C_i/C_f is equal to 2^{i-1-N} for the BWCDAC, while it is 1 for the FGDAC.

Table 4 summarizes the ratios, a_1, a_2, b_1, b_2, c_1, and c_2 for different values of load, amplifier, and unit-capacitance values. In this table, a_1, b_1, and c_1 represent the ratios for one-stage amplifier case, while a_2, b_2, and c_2 represents the ratio for two-stage amplifier case. From this table, it can be observed that for large values of C_{amp} the performance of the FGDAC in terms of the noise contributions from the amplifier, the switches, and the references can be improved compared to the BWCDAC.

8.2 Circuit description of FGDAC

The FGDAC is designed to obtain a low-power and compact DAC that can be integrated with larger systems. It is composed of several sub-blocks including an operational amplifier, epots, a buffer, switches, and a serial shift register. While the design of the FGDAC is slightly different, it is functionally the same as the BWCDAC.

In this implementation, the serial shift register is utilized to load the FGDAC digital data. This digital input word controls the desired output voltage by switching the individual capacitors between the reference voltage and the corresponding epot output voltage. This operation results in a charge on the input capacitors, which is then amplified by the charge amplifier to produce a voltage that can be expressed as

$$V_{ref} - V_{out} = \frac{1}{C_f} \sum_{i=1}^{n} a_i C_i (V_i - V_{ref}) \tag{56}$$

Table 4. Ratio of noise contributions from switches, references, and amplifier. $G_m R_{on} = 1/x$ **and** $R_{on}C \cdot GB = 1/y$**.**

Capacitors	a_1	a_2	b_1	b_2	c_1	c_2
$C \gg C_L$ & $NC \gg C_{amp}$	$\frac{x}{2^N}$	-	$\frac{4}{2^N N}$	-	$\frac{1}{2^N}$	-
$C_L \gg 2^N C$ & $NC \gg C_{amp}$	$\frac{x}{2^N} \cdot \frac{C_L}{C}$	-	$\frac{2}{N}$	-	$\frac{1}{2}$	-
$C_{amp} \gg C_L$ & $C_L \gg 2^N C$	$\frac{x}{2^N N} \cdot \frac{C_L C_{amp}}{C^2}$	$\frac{y}{(2^N N)} \cdot \frac{C_{amp}}{C}$	$\frac{2^{N+2}}{N^2}$	$\frac{2^{N+2}}{(N+1)^2}$	$\frac{2^N}{N}$	$\frac{2^N}{N}$
$C_{amp} \gg 2^N C$ & $C \gg C_L$	$\frac{x}{(2^N N)} \cdot \frac{C_{amp}}{C}$	$\frac{y}{(2^N N)} \cdot \frac{C_{amp}}{C}$	$\frac{4}{N^2}$	$\frac{2^{N+2}}{(N+1)^2}$	$\frac{1}{N}$	$\frac{2^N}{N}$
$C \gg C_{amp}$	-	$\frac{y}{2^N}$	-	$\frac{4}{(N+1)^2}$	-	$\frac{1}{2}$

where V_{ref} is the reference voltage, V_i is the epot output voltage, C_f is the feedback capacitor, C_i is the input capacitor, and a_i is the digital input bit for $i = 1, 2, ..., N$. In this implementation, equal size input/feedback capacitors are used.

The epots are used to set the scaled input voltages in (56). The block diagram of the epot is shown in Figure 13a, and is a modified version of the epot presented in [73]. This modified voltage reference is composed of a low-noise amplifier integrated with floating-gate transistors and programming circuitry that enables the tuning of the stored analog voltage. The amplifier, illustrated in Figure 13b, in the epot structure is used to buffer the stored analog voltage, enabling the epot to achieve low noise, low output resistance, as well as the desired output voltage range. 10 epots storing the scaled voltages are used to implement a $10-bit$ DAC. During programming the epots are controlled and read by employing a decoder.

In this architecture, epots and inverting amplifiers are the main blocks that use floating-gate transistors to exploit their analog storage and capacitive coupling properties. The epots employ floating-gate transistors to store the analog voltages, and the inverting amplifier uses them for their capacitive coupling properties and for removing the offset at its floating-gate terminal. A precise tuning of the stored voltage on floating-gate nodes is achieved by utilizing the hot-electron injection and the Fowler-Nordheim tunnelling mechanisms.

In this DAC implementation, no layout technique is employed for the input capacitor array. As expected, due to inevitable mismatches between the capacitors, there will be a gain error contributed from each input capacitor when epots are programmed without taking these mismatches into account. Therefore, after the initial epot programming, the stored voltages are also trimmed to compensate for these mismatches. The stored epot voltage is tuned by changing the floating-gate charge through the use of the internal programming circuitry. Programming of the epots is controlled via digital signals, *select*, *tunnel*, and *inject*. This digital control of the epot programming allows for the epot voltage to be adjusted to within $100\mu V$ of the desired voltage.

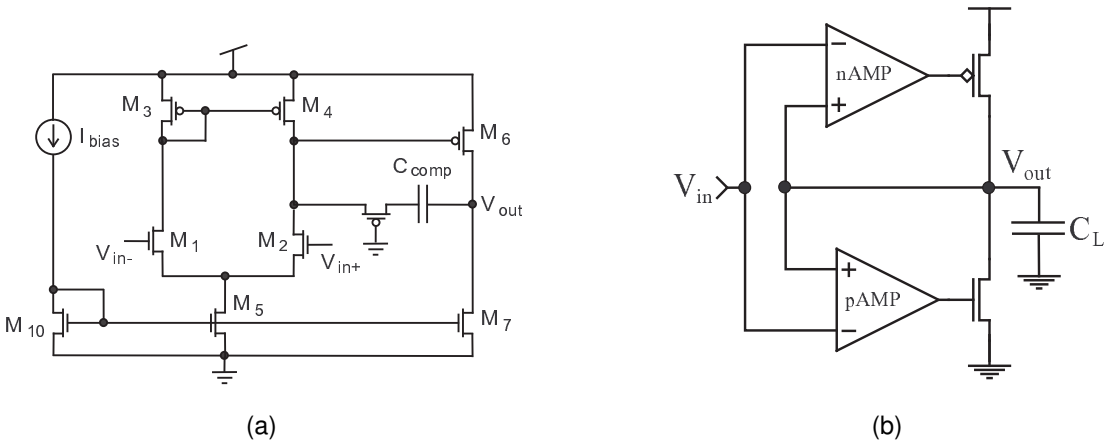


Figure 50. (a) Inverting amplifier schematic. I_{bias} is the bias current and C_{comp} is the compensation capacitor of the amplifier. (b) Implemented buffer using a push-pull output stage to drive the DAC output signal off-chip. $nAMP$ and $pAMP$ are the $nFET$ and $pFET$ input single-stage amplifiers, and C_L is the load capacitor.

Epots are required to drive capacitive loads when integrated into the FGDAC. Depending on the power consumption requirement, the output resistance of the epot amplifier can be set to allow operation at different converter speeds. The output resistance of the epot can be expressed as

$$R_{out} = \frac{R_{II}}{1 + g_{m_2} g_{m_6} R_I R_{II}} \tag{57}$$

where g_{m_2} and g_{m_6} are the transconductance of M_2 and M_6, and R_I and R_{II} are the output resistance of the first and second stages, respectively. Here, R_I is approximately equal to the output resistance of M_4, and R_{II} is the parallel combination of the output resistances of M_6 and M_7.

The inverting amplifier of the FGDAC is a two-stage amplifier as shown in Figure 50a. The FGDAC implementation with one-stage amplifier is described in [75]. The two-stage amplifier circuit allows to obtain a high gain and a large output swing [76]. The charge on the floating-gate node of this amplifier is precisely programmed by monitoring the amplifier output while the system operates in the reset mode. In this mode, all the input voltages to the input capacitors are set to the reference voltage. This condition ensures that the amplifier output voltage becomes equal to the reference voltage when the charge on its

floating-gate terminal is compensated. For this purpose, a pFET and a tunnelling junction are integrated with the floating-gate terminal of the amplifier for injection and tunnelling, respectively. By using this technique, the offset of the amplifier is reduced to much less than $1mV$. Lastly, a negative-feedback output stage [77], shown in Figure 50b, is employed to be able to buffer the output voltage off-chip. This buffer uses complementary single-stage error amplifiers for its shunt negative-feedback to achieve low-output resistance.

8.3 Measurement Results

In this section, we present the experimental results from the FGDAC architecture that was fabricated in a $0.5\mu m$ CMOS process. The previous results from the FGDAC with one-stage amplifier was presented in [75]. For the static and dynamic tests, the input data of the FGDAC is loaded using an on-chip serial shift register.

The input-output characteristic of the FGDAC is shown in Figure 51a. Epots are programmed to obtain $3V$ output voltage range with $LSB = \pm 1.5mV$. The integral and differential non-linearity (INL and DNL) of the FGDAC is tested with a static input using an all-codes test. From these tests, INL and DNL are found as shown in Figure 51b, and 51c, respectively. INL is limited between $0.35LSB$ to $-0.3LSB$, while DNL is measured to be between $0.35LSB$ to $-0.3LSB$. Within the full-scale range, the FGDAC yields better than $10-bit$ of static linearity. In these experiments, the static linearity of the FGDAC is mainly limited by the noise in the experimental set-up. The epot voltages are programmed with a resolution of $100\mu V$; higher DAC linearity would require tighter programming resolution as well as lower DAC noise levels. High resolution of the epots makes this implementation realizable for higher DAC resolutions. Also, flicker noise in the signal path was another limiting factor for the static measurements. Therefore, the DAC amplifier as well as the buffer need to be designed for low flicker noise to achieve a better DAC voltage trimming.

For the transient measurements, the digital data is loaded into the shift register at $3.4MHz$ clock frequency for a $170kHz$ sampling frequency. Dynamic measurements of

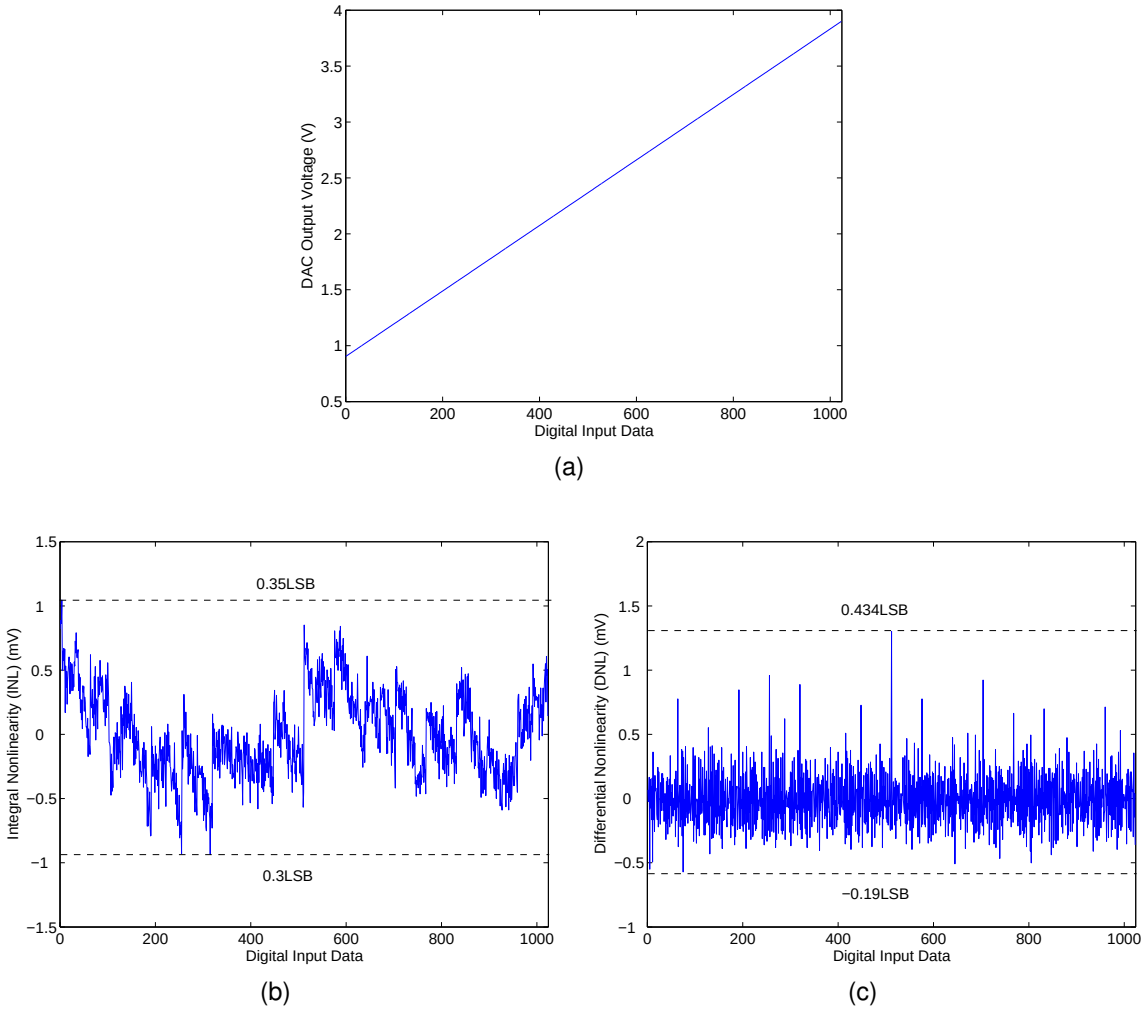


Figure 51. Experimental results obtained to characterize the static behavior of the $10-bit$ **FGDAC. (a) Output response of the FGDAC to** $10-bit$ **digital input code. The voltage output is a linear function of the digital input word. (b) INL characterization results for** $10-bit$ **digital input code. (c) DNL measurements of the FGDAC.**

the FGDAC are obtained by testing the performance of the DAC for 95% of a full-scale sinusoidal signal, as shown in Figure 52a. Also, the power spectrum of the output signal is shown in Figure 52b. It is observed that the FGDAC yields an $SFDR$ of $63.3dB$ for $1kHz$ output signal.

In this design, the unit capacitor is sized as $300fF$. The area of the individual blocks are summarized in Table 5, and the die photo of the fabricated chip is shown in Figure 53. The total DAC area including all the blocks are is around $0.117mm^2$, and the total die area for the DAC including all the wires and blocks is $0.208mm^2$. If this DAC was implemented

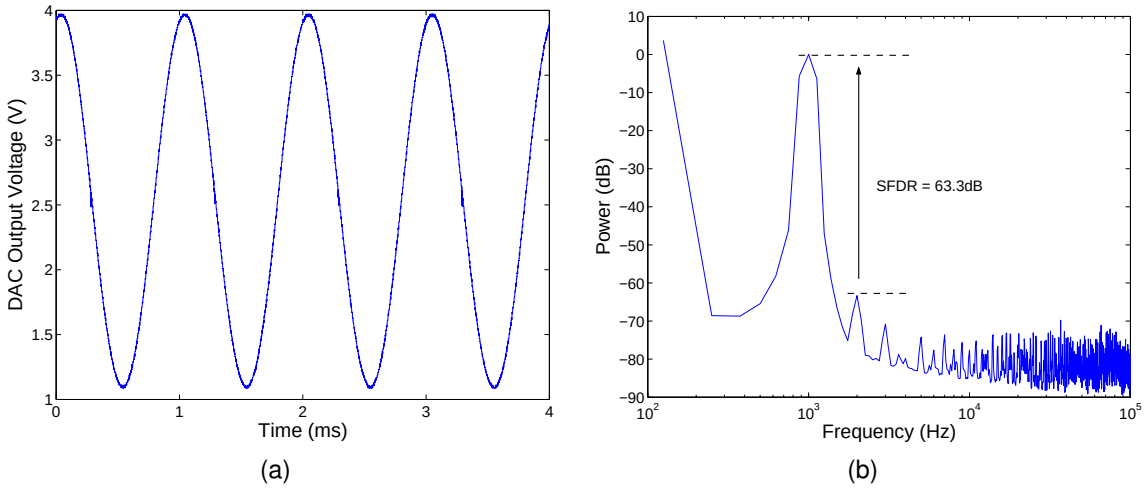


Figure 52. Dynamic measurements of the FGDAC: (a) $1kHz$ sinusoidal output response of the FGDAC. (b) Normalized power spectrum of $1kHz$ and $3.8V_{pp}$ signal created by the FGDAC.

by using a binary-weighted capacitor array, the total DAC area would be $0.644mm^2$ for the same size unit capacitor. Therefore, the $10-bit$ FGDAC yields around 3 times improvement in the total DAC area compared to the $10-bit$ BWCDAC. The parameters of the FGDAC based on the measurements and fabricated design is summarized in Table 6.

To illustrate the total design gain of the FGDAC relative to the BWCDAC, the design parameters are compared based on the assumption that the unit capacitor of the BWCDAC is 10 times smaller than the unit capacitor of the FGDAC. In addition, the amplifier and load capacitances are chosen as $C_{amp} = C_u$ and $C_L = 10C_u$, where C_u is the unit capacitance of the FGDAC. The results are summarized in Table 7. It is observed that when designed with one-stage amplifier the FGDAC operates around 10 times faster than the BWCDAC, and occupies 2 times smaller than the BWCDAC. In the area calculation, it is assumed that BWCDAC does not employ any layout technique, but in reality BWCDAC has to employ

Table 5. Area used for the FGDAC and its components.

Decoder	Epots	Capacitor area	DAC amplifier
$17{,}356\mu m^2$	$36{,}774\mu m^2$	$4{,}737\mu m^2$	$5{,}510\mu m^2$
Buffer	Biases	Shift register	Total DAC area
$10{,}962\mu m^2$	$22{,}134\mu m^2$	$20{,}100\mu m^2$	$208{,}073\mu m^2$

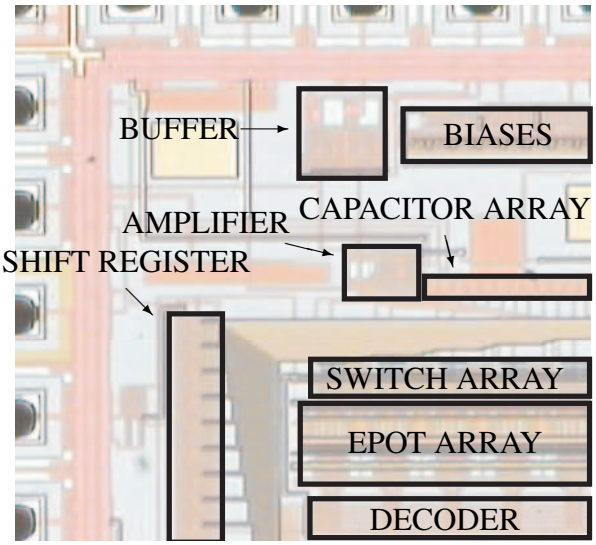


Figure 53. Die photo of the fabricated chip.

it to improve its linearity. Therefore, the gain in the capacitor area is assumed to much higher with the FGDAC design. The trade-off with the FGDAC design is that the amplifier contributes around 5 times ($\sqrt{b}$) more to the total DAC noise compared to the amplifier in the BWCDAC. As long as the amplifier noise is kept below the other noise sources, the FGDAC can provide better linearity with less area and faster speed.

In this chapter, an implementation of an epot-based floating-gate tunable DAC is described. Also, it is shown that it is a good candidate for implementing a compact and low-power DAC. This structure can be used for a wide range of embedded system applications where power and area become one of the main concerns. The results illustrate the flexibility and programmability of this architecture, which can be leveraged to create linear

Table 6. Parameters of the FGDAC.

Process	$0.5\mu m$ CMOS, 2 poly
Power supply	$5V$
Linearity (INL/DNL)	$>10-bit$
SFDR at 1kHz and 170Ksample/s	$63.3db$
Epot Programming Resolution	$100\mu V$
Programming Mechanisms	Hot-Electron Injection and Electron Tunnelling
Input capacitor	$300fF$
DAC area	$0.208\ mm^2$

Table 7. Design example for 10-bit DAC: Performance and area comparison. Unit Capacitors of BWC-DAC and FGDAC: $C = 0.1C_u$ and $C = C_u$. $C_{amp} = C_u$, $C_L = 10C_u$. Area: $A_{epot} = 10A_{C_u}$. $x = y = 100$.

Parameters	One-stage amplifier	Two-stage amplifier
a	12.8	1.17
b	$42 \cdot 10^{-3}$	$67 \cdot 10^{-3}$
c	0.42	0.6
$\tau_{BWCDAC}/\tau_{FGDAC}$	9.4	0.17
A_{BWCDAC}/A_{FGDAC}	1.84	1.84

or non-linear output voltage spacing. Dynamic re-calibration can also be achieved using this programmability feature to accommodate varying operating conditions.

CHAPTER 9

A RECONFIGURABLE MIXED-SIGNAL VLSI IMPLEMENTATION OF DISTRIBUTED ARITHMETIC

The battery lifetime of portable electronics has become a major design concern as more functionality is incorporated into these devices. Therefore, the shrinking power budget of modern portable devices requires the use of low-power circuits for signal processing applications. The data or media in these devices is generally stored in a digital format but the output is still synthesized as an analog signal. Examples of such devices are flash memory and hard disk based audio players. The signal processing functions employed in these devices include finite impulse response (FIR) filters, discrete cosine transforms (DCTs), and discrete Fourier transforms (DFTs). The common feature of these functions is that they are all based on the inner product. DSP implementations typically make use of multiply-and-accumulate (MAC) units for the calculation of these operations, and the computation time increases linearly as the length of the input vector grows. In contrast, distributed arithmetic (DA) is an efficient way to compute an inner product. It computes an inner product in a fixed number of cycles, which is determined by the precision of the input data. It has been employed for image coding, vector quantization, discrete cosine transform and adaptive filtering implementations [78–81].

DA is computationally more efficient than MAC-based approach when the input vector length is large. However, the trade-off for the computational efficiency is the increased power consumption and area usage due to the use of a large memory. These problems can be alleviated by utilizing mixed-signal circuit implementations for optimized DA performance, power consumption, and area usage. In this work, we propose a mixed-signal DA architecture built by utilizing the analog storage capabilities of floating-gate transistors for reconfigurability and programmability. The circuit compactness is obtained through the

application of the iterative nature of the DA computational framework, where many multipliers and adders are replaced with an addition stage, a single gain multiplication, and a coefficient array.

In this chapter, the computational efficiency of DA implementation is demonstrated by configuring it as an FIR filter. The low-power implementations of these filters can readily ease the power consumption requirements of portable devices. Also, due to the serial nature of the DA computation, the power and area of this filter increase linearly with its order. Hence, this design approach allows for a compact and low-power implementation of high-order FIR filters.

In the next section, the DA computation is described. Subsequently, the hybrid DA architecture is explained, and the integration of tunable voltage references into the DA implementation is described. After that, the experimental results of this reconfigurable DA for FIR filtering are presented. In the last part of the chapter, the characteristics of the proposed implementation is summarized.

9.1 DA computation

The DA concept was first introduced by Croisier et al. [82], and later utilized for the hardware implementation of digital filters using memory and adders instead of multipliers [83]. It is an efficient computational method for computing the inner product of two vectors in a bit-serial fashion [84]. The operation of DA can be derived from the inner product equation as follows

$$y[n] = \sum_{i=0}^{M-1} x[n-i]w[i] \tag{58}$$

In the case of FIR filtering, x is the input vector and w is the weight vector. Using a K-bit 2's-complement representation, x can be written as $x[n-i] = -b_{i0} + \sum_{j=1}^{K-1} b_{ij}2^{-j}$, where b_{i0} is the sign bit, b_{ij} is the j^{th} bit of the i^{th} element in the vector x, and $b_{i(K-1)}$ is the least significant bit. Substituting x into (58), and by reordering the summations and grouping

the terms together, (58) can be written as

$$y[n] = -\sum_{i=0}^{M-1} w_i b_{i0} + \sum_{j=1}^{K-1} 2^{-j} \sum_{i=0}^{M-1} w_i b_{ij} \tag{59}$$

In digital implementations, the summation, $\sum_{i=0}^{M-1} w_i b_{ij}$, is pre-computed and stored in a memory for multiplier-less operation and reduced hardware complexity. This is usually achieved by storing 2^M possible combinations of summed weights in the memory, which simplifies the hardware requirements of DA to a bank of input registers, a memory, a delay element, a shifter, a switch, and an adder as illustrated by Figure 54a. By reusing the hardware K times, an output sample can be processed in K clock cycles regardless of the number of taps, M, and without using a multiplier. Digital DA architectures obtain significant throughput advantages when M is large.

In contrast to digital implementations, the addition in the analog domain is much more power and area efficient. Therefore, the high memory usage of digital DA implementations can be eliminated by processing the digital input data in the analog domain as shown in Figure 54b. To design such a structure, weights in (58) are stored in the analog domain. For an individual weight, data is processed in a similar way as it is achieved by serial DACs, where the conversion is performed sequentially.

9.2 Proposed DA architecture

The hybrid DA architecture consists of four components, which are a 16-bit shift register, an array of tunable FG voltage references (*epots*) [73], inverting amplifiers (*AMP*), and sample-and-hold (SH) circuits, as illustrated in Figure 55. The timing of the digital data and control bits governs the DA computation and is illustrated in Figure 56. Digital inputs are introduced to the system by using a serial shift register. These digital input words represent the digital bits, b_{ij} in (59), which selects the epot voltages to form the appropriate sum of weights necessary for the DA computation at the j^{th} bit. The clock frequency of the shift register is dependent on the input data precision, K, and the length of the filter, M, and is equal to $M \cdot K$ times the sampling frequency. Once the j^{th} input word is serially loaded

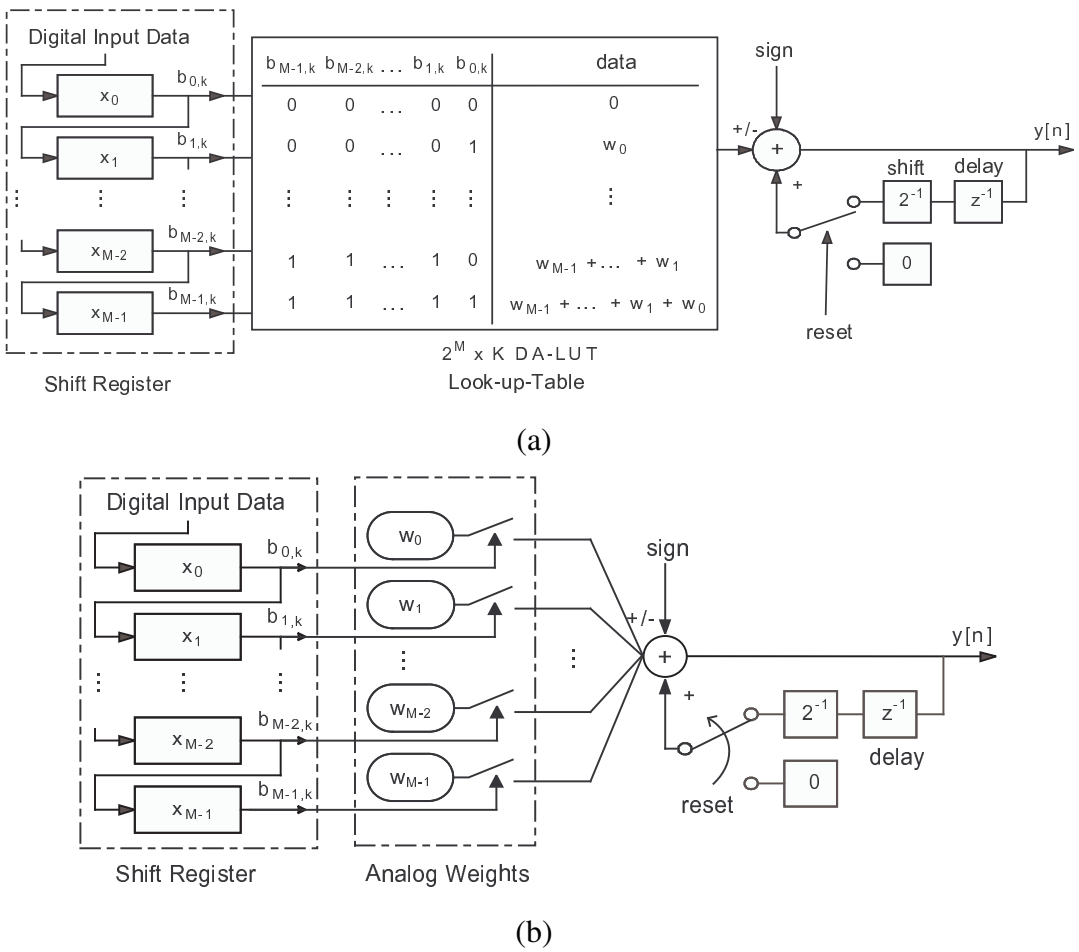


Figure 54. Basic DA hardware architecture. $b_{i,k}$ is the input bit for k^{th} cycle of operation and $y[n]$ is the output. (a) Digital implementation. (b) Proposed hybrid mixed-signal implementation using digital input data and stored analog weights. Digital input data is processed in the analog domain.

into the top shift register, the data from this register is latched at K times the sampling frequency. If the area used by the shift registers is not a design concern, then ideally an M-tap FIR filter should have M shift registers. A clock that is K times faster than the sampling frequency would be used for this ideal configuration.

The analog weights of DA are stored by the epots. When selected, these weights are added by employing a charge amplifier structure composed of same size capacitors, and a two-stage amplifier, AMP_1. The epot voltages as well as the rest of the analog voltages in the system are referenced to a reference voltage, $V_{ref} = 2.5V$. Since the addition operation is performed by using an inverting amplifier, the relative output voltage, when *Reset* signal

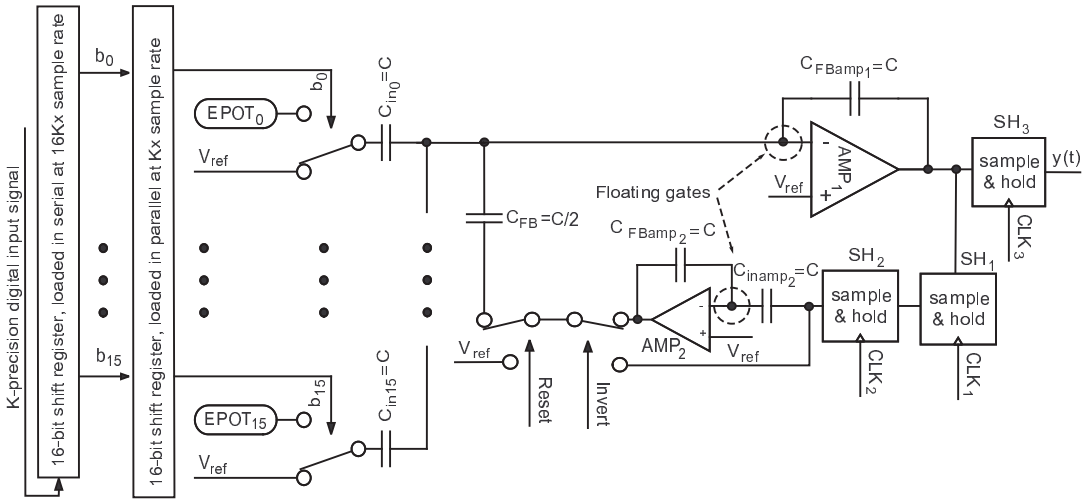


Figure 55. Implementation of the 16-tap hybrid FIR filter. b_i is the input bit for j^{th} cycle of operation and $y(t)$ is the output. Epots store the analog weights. Sample-and-holds, SHs, are used to obtain the delay and hold the computed output voltage.

is enabled, becomes equal to the negative sum of the selected weights for $C_{in_i} = C_{FBamp_1}$. For the first computational cycle, the result of the addition stage represents the summation, $\sum_{i=0}^{m-1} w_i b_{i(K-1)}$ in (59), which is the addition of weights for the LSBs of the digital input data.

In the feedback path of the system, a delay, an invert and a divide-by-two operations are used for the DA computation. For that purpose, sample-and-hold circuits, SH_1 and SH_2, and inverting amplifiers, AMP_1 and AMP_2, are employed in the implementation. The SH circuits store the amplifier output to feed it back to the system for the next cycle of the computation. Non-overlapping clocks, CLK_1 and CLK_2, are used to hold the analog voltage while the next stream of digital data is introduced to the addition stage. These clocks have a frequency of K times the sampling frequency. The stored data is then inverted relative to the reference voltage by using the second inverting amplifier, AMP_2, to obtain the same sign as the summed epot voltages. AMP_2 is identical to AMP_1, and has the same size input/feedback capacitors. After obtaining the delay and the sign correction, the stored analog data is fed back to the addition stage as delayed analog data. During the addition, it is also divided by two by using $C_{FB} = C_{FBamp1}/2 = C/2$, which gives a gain of 0.5 when it

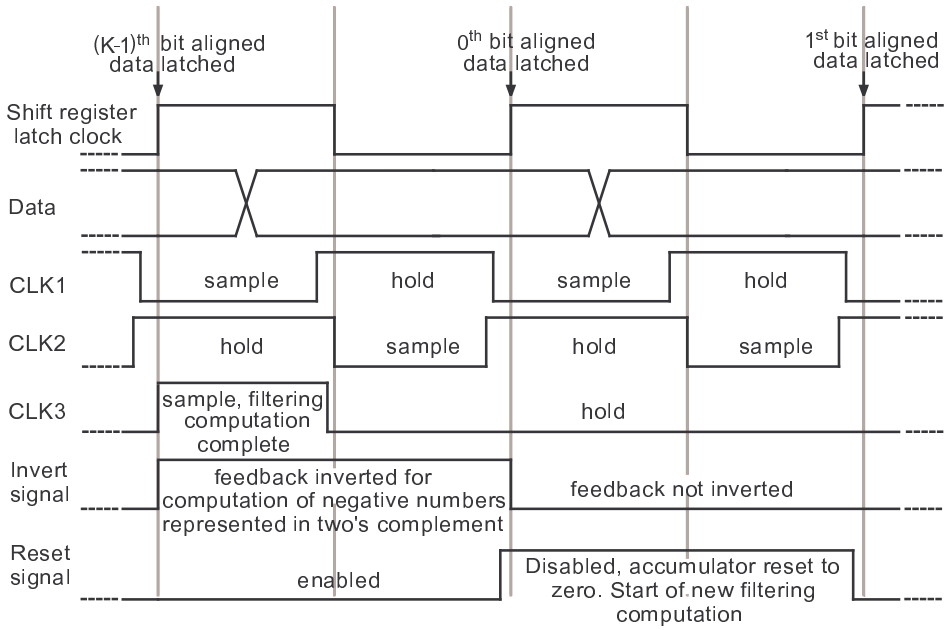


Figure 56. Digital clock diagram of the filter architecture. For desired sampling frequency, f_s, $K-bit$ precise $M-bit$ digital input data is loaded serially to a shift register at a $K \cdot M \cdot f_s$ clock frequency, and latched at a $K \cdot f_s$ clock frequency. CLK_1, CLK_2, and CLK_3 are the bits used to control SH_1, SH_2, and SH_3, respectively. *Invert* signal is used to obtain 2's-complement compatibility. Also, *Reset* signal is used to clear the result of the previous computation.

is added to the new sum. This operation is repeated until the MSBs of the digital input data is loaded into the shift register. The MSBs correspond to $(K-1)^{th}$ bits, and are used to make the computation 2's-complement compatible. This compatibility is achieved by disabling the inverting amplifier in the feedback path during the last cycle of the computation by enabling the *Invert* signal. As a result, during the last cycle of the computation, the relative output voltage of AMP_1 becomes

$$V_{out_{amp1}} - V_{ref} = -\sum_{i=0}^{M-1} \frac{C_{in_i}}{C_{FBamp_1}}(V_{ref} - V_{epot_i})b_{i0} + \sum_{j=1}^{K-1} 2^{-j} \sum_{i=0}^{M-1} \frac{C_{in_i}}{C_{FBamp_1}}(V_{ref} - V_{epot_i})b_{ij} \quad (60)$$

where the first term is the result of the calculation with the sign bits. Finally, when the computation of the output voltage in (60) is finished, it is sampled by SH_3 using CLK_3, which is enabled once every K cycle. SH_3 holds the computed voltage till the next analog output voltage is ready. The new computation starts by enabling the *Reset* signal to zero out the effect of the previous computation. Then, the same processing steps are repeated for the next digital input data.

9.3 Circuit description of computational blocks

To achieve an accurate computation using DA, the circuit components are designed to minimize the gain and offset errors in the signal path. In this architecture, those components are the epots, the inverting amplifiers, and the sample-and-holds.

The epot, shown in Figure 13a, is modified from its original version [73] to obtain a low-noise voltage reference. It is a dynamically reprogrammable, on-chip voltage reference that uses a low-noise amplifier integrated with floating-gate transistors and programming circuitry to tune the stored analog voltage. The amplifier in the epot circuit is used to buffer the stored analog voltage so that the epot can achieve low noise and low output resistance as well as the desired output voltage range. An array of epots is used for storing the filter weights; and during the programming, individual epots are controlled and read by employing a decoder.

In this architecture, epots and inverting amplifiers are the main blocks that use FG transistors to exploit their analog storage and capacitive coupling properties. A precise tuning of the stored voltage on FG node is achieved by utilizing the hot-electron injection and the Fowler-Nordheim tunnelling mechanisms. The epots employ FG transistors to store the analog coefficients of the inner product. In contrast, the inverting amplifiers use them not only to obtain capacitive coupling at their inverting-node, but also to remove the offset at their FG terminals.

One of the main advantages of exploiting FG transistors in this design is that the area allocated for the capacitors can be dramatically reduced. It is shown in [75] that epots can be utilized to implement a compact programmable charge amplifier DAC. This structure helps to overcome the area overhead, which is mainly due to layout techniques used to minimize the mismatches between the input and feedback capacitors. Similarly in this DA implementation, the unit capacitor, C, is set to $300fF$, and no layout technique is employed. As expected, due to inevitable mismatches between the capacitors, there will be a gain error contributed from each input capacitor. The stored weights are also used to compensate this

mismatch. When the analog weights are stored to the epots, the gain errors are also taken into account to achieve accurate DA computation.

Unlike switched-capacitor amplifiers, the addition in this implementation is achieved without resetting the inverting node of the amplifiers. This is because the floating-gate inverting-node of the amplifiers allow for the continuous-time operation. This design approach eliminates the need for multi-phase clocking or resetting. Inverting amplifiers are implemented by using a two-stage amplifier structure [76], shown in Figure 57a, to obtain a high gain and a large output swing. Similar to the epots, the charge on the floating-gate node of these amplifiers is precisely programmed by monitoring the amplifier output while the system operates in the reset mode. In this mode, the shift registers are cleared and the *Reset* signal is enabled. Therefore, all the input voltages to the input capacitors including the voltage to the feedback capacitor, C_{FB}, are set to the reference voltage. These conditions ensure that the amplifier output becomes equal to the reference voltage when the charge on the floating-gate is compensated. The charge on the floating-gate terminal is tuned using the hot-electron injection and the Fowler-Nordheim tunnelling mechanisms. By using this technique, the offset at the amplifier output is reduced to less than $1mV$.

Lastly, SH circuits need to be designed to simultaneously achieve high sampling speed and high sampling precision due to the bit-serial nature of the DA computation. Therefore, these circuits are implemented by utilizing the sample-and-hold technique using Miller hold capacitance [85], as illustrated in Figure 57b. This compact circuit minimizes the signal dependent error, while maintaining the sampling speed and precision by using the Miller capacitance technique together with Amp_3 shown in Figure 57c. For simplification, if we assume there is no coupling between M_1 and M_2, and amplifier, Amp_3, has a large gain, then the pedestal error contributed from turning switches (M_1 and M_2) off can be written as

$$\Delta V_{S1} + \Delta V_{S2} = \frac{\Delta Q_1(C_2 + C_{2B})}{C_{2B}(C_1 + C_2) + C_1C_2(A + 1)} + \frac{\Delta Q_2}{C_2} \tag{61}$$

where ΔQ_1 and ΔQ_2 are the charges injected by M_1 and M_2, respectively. Also, A and

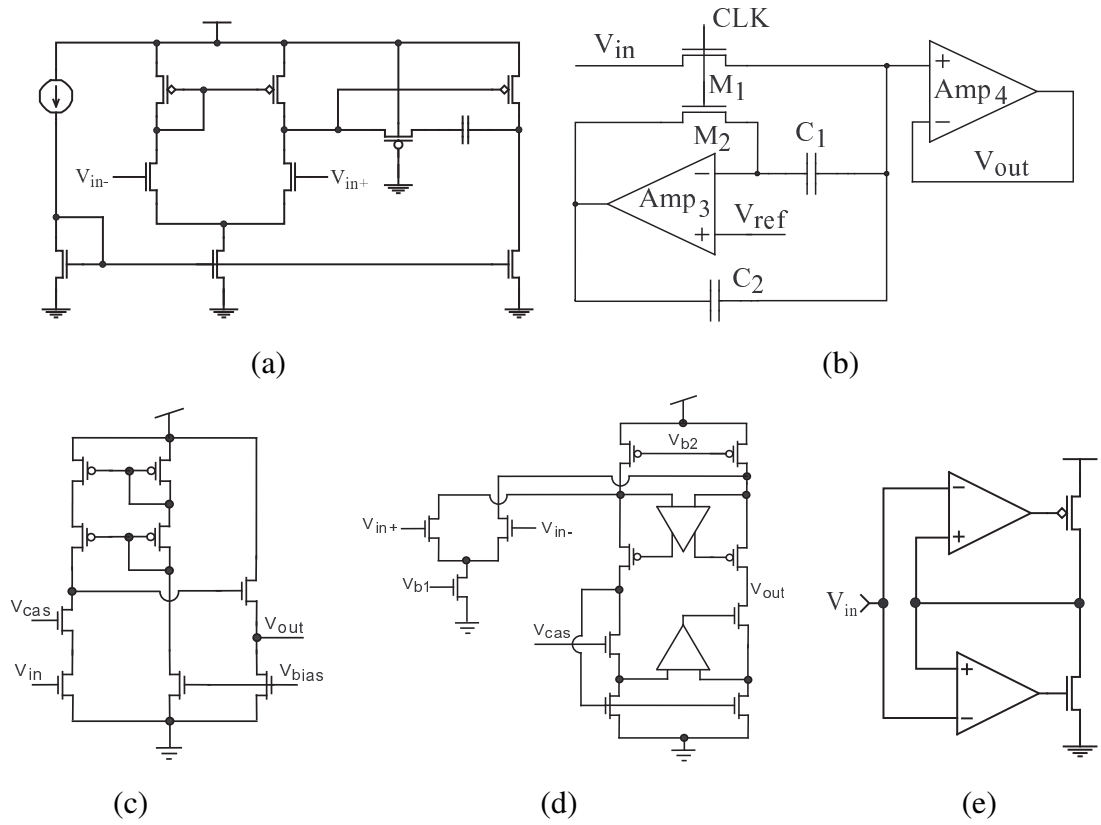


Figure 57. Circuit components. (a) Inverting amplifier schematic. This circuit is used for AMP_1 and AMP_2 in the DA implementation. (b) Sample and hold circuit schematic. This circuit is employed for SH_1 and SH_2 in the DA implementation. (c) Amp_3 in the sample and hold circuit. (d) Amp_4 in the sample and hold circuit. (e) Buffer schematic. This circuit is used to drive the sinal off-chip.

C_{2B} are the gain and input capacitance of the amplifier, Amp_3. ΔQ_2 is independent of the input level, therefore ΔV_{S2} can treated as an offset. In addition, the error contributed by M_1, ΔV_{S1}, can be minimized by the Miller feedback, and this error decreases as A increases [85]. Due to serial nature of the DA computation offset in the feedback path is attenuated as the precision of the digital input data increases. Therefore, Amp_3 is designed to minimize mainly the signal dependent error, ΔV_{S1}.

Moreover, a gain-boosting technique [86] is incorporated into the SH amplifier, Amp_4, as shown in Figure 57d, to achieve a high gain and fast settling. Two SH circuits are used in the feedback path to obtain the fixed delay for the sampled analog voltage. In addition, the third SH is utilized to sample and hold the final computed output once every K cycles. This

SH uses a negative-feedback output stage [77], shown in Figure 57e, to be able to buffer the output voltage off-chip. Due to the performance requirements of the system, these SH circuits consume more power than the rest of the system.

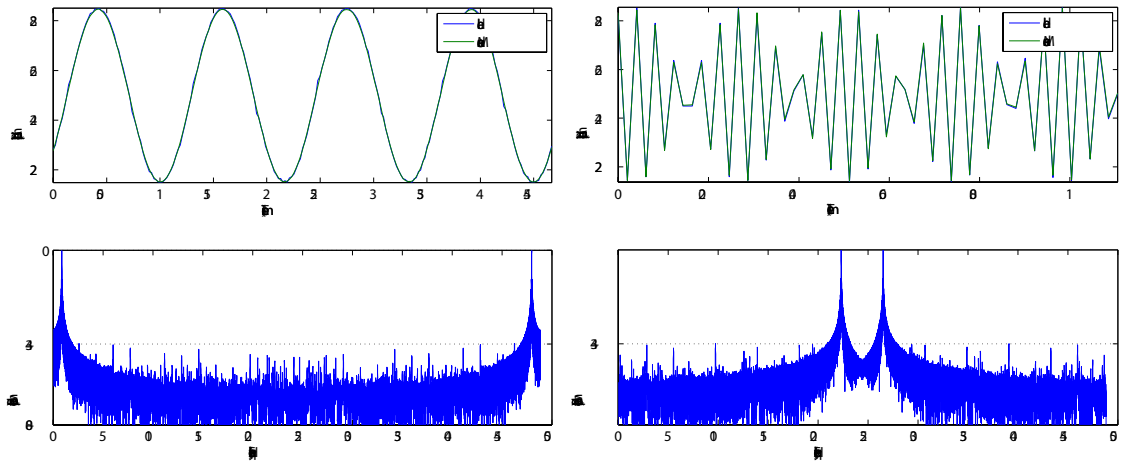
9.4 Measurement Results

In this section, we present the experimental results from the proposed DA architecture, which is configured as an FIR filter. The measurement results are obtained from the chips that were fabricated in a $0.5\mu m$ CMOS process. This 16-tap FIR filter is designed to run at $32/50kHz$ sampling frequency depending on the desired performance. The precision of the digital input data is set to 8 for these experiments. To meet this sampling rate, the data is loaded into the upper shift register at a rate of $3.84MHz$ for a $32kHz$ sampling frequency or $6.4MHz$ for a $32kHz$ sampling frequency.

To demonstrate the reconfigurability, the filter is configured as a comb, a low-pass, and a band-pass filter. The coefficients of these filters are shown in Table 8. Ideal coefficients are given to illustrate how close the epots are programmed to obtain the actual coefficients. The epots are programmed relative to a reference voltage, V_{ref}, which is set to $2.5V$. The error of the stored epot voltages are kept below $1mV$ to minimize the effect of weight errors on the filter characteristics.

An $858Hz$ sinusoidal output of the low-pass filter at a $50kHz$ sampling rate is illustrated in Figure 58a. The spurious-free-dynamic-range (SFDR) of this signal is measured to be $43dB$. For the comb filter with a $22kHz$ input signal frequency, it is observed that the SFDR does not degrade as shown in Figure 58b. Although the input precision was set to 8 bits, the gain error in the system as well as noise in the experimental set-up limits the maximum achievable SFDR.

The second experiment is performed to characterize the magnitude and phase responses of the filters. For that purpose, a sinusoidal wave at a fixed sampling rate, $32/50kHz$, is generated using the digital data, and the magnitude and phase responses are measured by



(a) (b)

Figure 58. Transient responses for $50kHz$ sampling frequency and their power spectrums. (a) Low-pass filter output has a frequency of $858Hz$. (b) Comb filter output has a frequency of $22kHz$.

sweeping the frequency of the input sine wave from DC to $16/25kHz$. For this experiment, 256 data points are collected to accurately measure the frequency response of these filters. These responses follow the ideal responses closely even if the sampling rate is increased as illustrated in Figures 59a, 59b, and 59c. Any variation in the frequency response as the sampling rate increases is caused by the noise and offset in the feedback path as well as due to the performance degradation of the circuits. As the output signal amplitude becomes very low, the experimental set-up limits the resolvable magnitude and phase. As expected for a symmetrical FIR filter, the measured phase responses of comb, low-pass, and band-pass filters are linear.

The static power consumption of the fabricated chip is measured as $16mW$. Most of the power is consumed by the SH and inverting amplifier circuits. The die photo of the designed chip is shown in Figure 60. The system occupies around half of the $1.5 \cdot 1.5mm^2$ die area. The cost to increase the filter order is $0.011mm^2$ of die area and $0.02mW$ of power for each additional filter tap. This readily allows for the implementation of high-order filters. Lastly, the performance of the filter is summarized in Table 9.

Table 8. Ideal and actual (programmed epot voltages) coefficients of the comb, low-pass, and band-pass filters.

Filter	Comb		LPF		BPF	
Coefficients	Ideal	Actual (V)	Ideal	Actual (V)	Ideal	Actual (V)
1	0.4	2.0996	-0.0190	2.5192	0.033	2.4670
2	0	2.4994	-0.0390	2.5393	-0.064	2.5639
3	0	2.4994	0.0260	2.4738	-0.053	2.5530
4	0	2.5007	0.0160	2.4835	0.038	2.4617
5	0	2.5005	-0.0240	2.5239	0.047	2.4528
6	0	2.5000	-0.0360	2.5362	-0.054	2.5541
7	0	2.4999	0.0600	2.4401	-0.056	2.5561
8	0	2.4994	0.1800	2.3201	0.057	2.4425
9	0	2.4997	0.1800	2.3201	0.057	2.4427
10	0	2.5002	0.0600	2.4391	-0.056	2.5560
11	0	2.4998	-0.0360	2.5358	-0.054	2.5535
12	0	2.5002	-0.0240	2.5240	0.047	2.4527
13	0	2.5001	0.0160	2.4853	0.038	2.4616
14	0	2.5001	0.0260	2.4743	-0.053	2.5526
15	0	2.4997	-0.0390	2.5389	-0.064	2.5638
16	0.4	2.0996	-0.0190	2.5184	0.033	2.4669

Table 9. Performance and design parameters of the DA based FIR filter.

Process	$0.5\mu m$, $2-poly$ CMOS
Power supply	$5V$
Reference voltage	$2.5V$
Epot Programming Resolution	100μV
Programming Mechanisms	Hot-Electron Injection and Electron Tunneling
Unit capacitor	$300fF$
Sampling frequency	$30/50KHz$
Input data precision	8
Number of filter taps	16
Increase in the power per tab	$0.02mW$
Increase in the area per tab	$0.011mm^2$
Total static power consumption	$16mW$
Used chip area	$\sim 1.125mm^2$

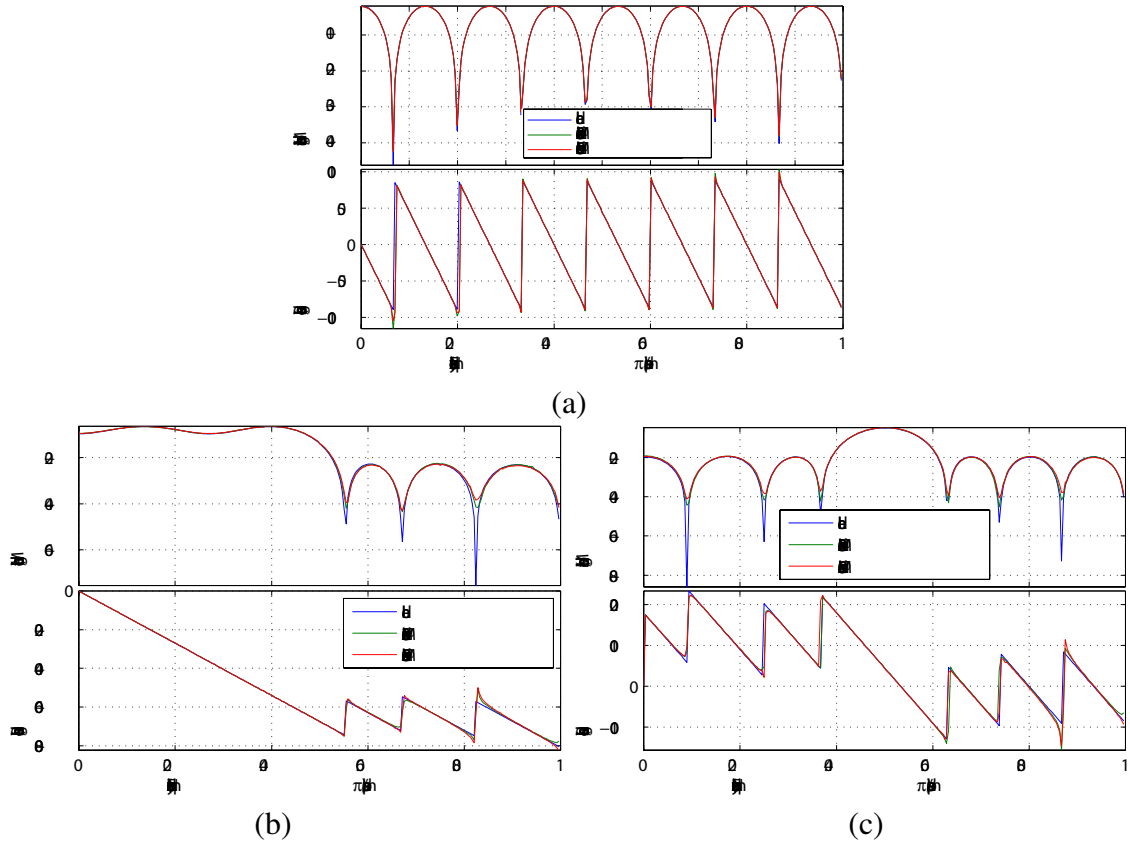


Figure 59. Magnitude and phase responses at $32/50kHz$ sampling rates. (a) Comb filter. (b) Low-pass filter. (c) Band-pass filter.

9.5 Discussion

The proposed DA structure which can be used for FIR filtering circumvents some of these problems by employing DA for signal processing and utilizing the analog storage capabilities of floating-gate transistors to obtain programmable analog coefficients for reconfigurability. In this way, the DAC is used as a part of the DA implementation, which helps achieving digital-to-analog conversion and signal processing at the same time.

Compared to the switched-capacitor implementations, which have their coefficients set by using different capacitor ratios, the proposed implementation offer more design flexibility since its coefficients can be set by tuning the stored weights at the epots. Also, offset accumulation and signal attenuation make it difficult to implement long tapped delay lines with these approaches. In the proposed implementation, we showed that DA processing

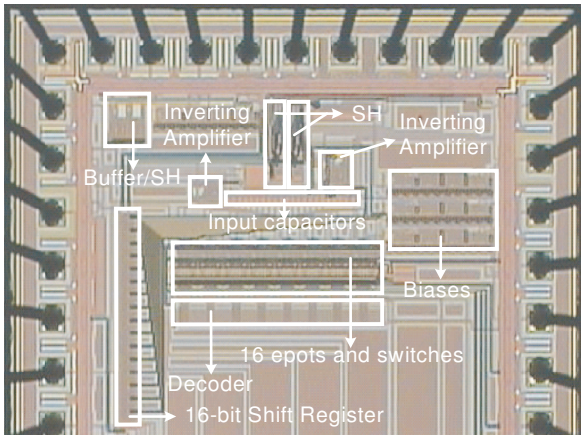


Figure 60. Die photo of the DA based FIR filter chip.

decreases the offset as the precision of the digital input data increases. Also, the gain error in this implementation is mainly caused by the two inverting stages (implemented using AMP_1 and AMP_2), and can be minimized using special layout techniques only at these stages. The measurement results illustrated that the output signal of the filter follows the ideal response very closely. This is mainly because it is mostly insensitive to the number of filter taps and most of the computation is performed in the feedback path. Also, the power and area of the proposed design increases linearly with the number of taps due to the serial nature of the DA computation. Therefore, this design approach is well suited for compact and low-power implementations of high-order filters for post-processing applications. The programmable analog coefficients of this filter will enable the implementation of adaptive systems that can be used in applications such as adaptive noise cancellation and adaptive equalization. Since DA is an efficient computation of an inner product, this architecture can also be utilized for signal processing transforms such as a modified discrete cosine transform.

CHAPTER 10

IMPACTS AND APPLICATIONS OF THE PRESENTED WORK

10.1 Impacts

In this work, a tunable voltage reference and a family of tunable resistors are designed to leverage the tunability and reconfigurability into the analog and mixed signal circuits. In this way, precision, accuracy, compactness, and power consumption issues associated with the technology scaling and digital circuit implementations are aimed to be alleviated. Therefore, the impact of the presented work can be summarized as below:

1. **Tunable CMOS resistor implementation using gate linearization technique:**

   I have designed, simulated, and tested a tunable floating CMOS resistor using floating-gate transistors and gate linearization technique. I also analyzed this technique to determine the limitations of the method. This resistor uses only 2 capacitors and 1 transistor in addition to the programming circuit. Therefore, depending on the desired performance the die area of this resistor can be easily optimized. Within the existing implementations of this technique, the proposed implementation does not consume additional power for the linearization and its operation does not depend on the supply rails. In addition, since floating-gate transistors can store analog voltages, this resistor stores its own resistor value, thus becomes very suitable for applications where an array of resistors are needed. It yields around 1.3% linearity for $1V_{pp}$ sinusoidal signals and its linearity is mainly limited by the body effect.

2. **Compact and tunable CMOS resistor using scaled-gate linearization technique:**

   I have designed, simulated, and tested a compact tunable CMOS resistor by employing the scaled-gate linearization technique. Although the resistor implementation based on the gate linearization is a floating resistor, by scarifying this feature and

operating it as a grounded resistor the tunable resistor using scaled-gate linearization yields more compact and more linear resistor. Better than $7-bit$ linearity is obtained for $1V_{pp}$ sinusoidal input. I have showed that the resistance and temperature coefficient of this resistor can be tuned to desired operating point. By using the stress tests, I have demonstrated that the resistance of this resistor drifts negligibly over time. Based on the worst-case data, it is calculated that the resistance drifts $1.6 \cdot 10^{-3}\%$ over the period of 10 years at 25^oC. Similar to the resistor circuit using gate-linearization, this circuit does not consume additional power for the offset generation and feedback computation.

I have demonstrated the use of this circuit by employing it in a data converter. I have designed, implemented, and tested a binary-weighted resistor DAC using this tunable floating-gate CMOS resistor. The software code to test this converter has been written by me and Mr. Haw-Jing Lo. I have demonstrated that $15-bit$ accurate $4-bit$ resolution DAC can be built using these resistors.

3. **Tunable Highly Linear Floating-Gate CMOS Resistor Using Common-mode Linearization Technique:**

   I have designed, simulated, and tested a tunable highly linear CMOS resistor by employing the common-mode linearization technique and floating-gate transistors. I have showed that this resistor offers a compact and power efficient implementation that yields around $72dB$ of linearity. I have analyzed the common-mode linearization method and showed the linearity limitations. I have also demonstrated the limitations in the implementation and showed the possible causes and their effects.

   I have employed this tunable resistor in the design of highly linear amplifier and two-quadrant multiplier circuits. I have designed, simulated, and tested these highly linear amplifier and multiplier circuits. The amplifier exhibited 0.018% THD for $1V_{pp}$ differential input, and a linear input range of $2.5V_{pp}$. With this implementation

it is possible to set the gain of the amplifier accurately and precisely.

4. **Tunable voltage reference:**

   I have designed, simulated, and tested a tunable voltage reference (epot). This epot has been designed to store the scale factors of a binary-weighted capacitor DAC and the coefficients of a distributed-arithmetic based FIR filter. For that purpose, it is designed to drive a large capacitive load while providing a stable voltage by keeping the noise, temperature variation, and charge loss minimized. The measured thermal noise level of this voltage reference is $-120dB$, and its noise corner is measured to be around $4kHz$. Also, its temperature coefficient is measured to be around $16.2ppm/^oC$. For an array of 10 epots programmed to different voltages, the mean temperature coefficient is measured as $16ppm/^oC$ with a maximum variation of $20.8ppm/^oC$. Moreover, based on the stress test it is calculated that the stored epot voltage drifts $10^{-3}\%$ over the period of 10 years at 25^oC. I have analyzed the noise, temperature dependence, and retention of this reference to correlate with the measured data.

5. **Programmable Voltage-Output Digital-to-Analog Converter:**

   I have designed, simulated, and tested a programmable 10-bit linear digital-to-analog converter. I have analyzed its noise, speed, and area. Also, I have compared its performance with the performance of a binary-weighted capacitor digital-to-analog converter. I have shown that when the unit capacitor of the BWCDAC is 10 times smaller than the unit capacitor of the FGDAC and these converter are designed with one-stage amplifier, the FGDAC operates around 10 times faster than the BWCDAC, and occupies more than 2 times smaller area than the BWCDAC does. Also, I have shown that as long as the amplifier noise is kept smaller than the other noise sources, the FGDAC can provide better linearity with less area and faster speed. Therefore,

this structure will enable very compact and low-power implementation of digital-to-analog converters.

The idea of this structure was first proposed by Dr. Paul E. Hasler, and the test codes to characterize this converter have been written by me along with Mr. Christopher M. Twigg and Mr. Haw-Jing Lo.

6. **A Reconfigurable Mixed-Signal VLSI Implementation of Distributed Arithmetic:**

   I have designed and simulated a reconfigurable distributed arithmetic (DA) architecture. Along with Mr. Walter Huang, I have tested this architecture and demonstrated its functionality for FIR filters. Compared to existing FIR filter implementations, the proposed implementation offer a more design flexibility since its coefficients can be set by tuning the stored weights.

   Offset accumulation and signal attenuation in the traditional FIR filter implementations make it difficult to implement long tapped delay lines with these approaches. In the proposed implementation, we have showed that DA processing decreases the offset as the precision of the digital input data increases. Also, the power and area of the proposed design increases linearly with the number of taps due to the serial nature of the DA computation. Therefore, this design approach is well suited for compact and low-power implementations of high-order filters for post-processing applications.

   The idea of this structure was first proposed by Dr. Paul E. Hasler, Dr. David Anderson, and Mr. Walter Huang.

10.2 Applications

The presented circuits can be used in a variety of applications where the accuracy and precision or the area and power consumption become the main design concerns. Some of these applications are listed below:

10.2.1 Tunable resistors

Due to their compactness and power efficiency, these resistors can be used in low-power implementations of the ANN systems for storing and tuning the weights.

Moreover, the resistor based on the common-mode linearization technique offers high linearity at the expense of very low power consumption. Therefore, as also demonstrated in this work, this resistor becomes very useful for highly linear amplifier and multiplier circuits. Also, it can be integrated in variable-gain-amplifier to set the gain of the amplifier to a desired point.

I have shown that in addition to the resistance of these resistors, their temperature coefficients can also be tuned. Therefore, they can be used in current reference and voltage reference circuits to implement tunable references with very low temperature coefficients.

Lastly, as demonstrated in this work, these resistors can be employed to implement digital-to-analog converters. Therefore, these resistors will enable the design of multi-bit CMOS quantizers to be used in pipelined and over-sampling data converters.

10.2.2 Epot

In this work, I have demonstrated that a tunable reference can built by using the analog storage capability of the floating-gate transistors. Also, I have showed that this tunable reference can be used to build a digital-to-analog converter and a reconfigurable distributed-arithmetic based FIR filter. In addition to these applications, epots can be used in the mixed-signal implementations of the infinite-impulse-response filters and correlators.

10.2.3 Mixed-signal implementation of the distributed arithmetic

The programmable analog coefficients of the distributed-arithmetic based FIR filter will enable the implementation of adaptive systems that can be used in applications such as adaptive noise cancellation and adaptive equalization. Since distributed arithmetic is an efficient computation of an inner product, this architecture can also be utilized for signal processing transforms such as a modified discrete cosine transform. Also, this distributed-arithmetic

architecture can be employed for image coding, vector quantization, and adaptive filtering implementations.

APPENDIX A

LINEARITY ANALYSIS OF GATE AND COMMON-MODE LINEARIZATION TECHNIQUES

In order to analyze these nonlinearities, the drain current of an nMOS transistor in the strong inversion is expressed as

$$I_d = \frac{\mu C_{ox} W}{L}[f(v_g, v_d, v_s) - g(v_b, v_d, v_s)] \tag{62}$$

where μ is the carrier mobility, C_{ox} is the gate capacitance per unit area, W is the channel width, L is the channel length, and v_g, v_d, v_s, and v_b are the gate, drain, source, body voltages (referenced to the ground), respectively [59]. Similar to the drain current, the carrier mobility is also dependent on the terminal voltages and can be expressed in terms of f and g as follows

$$\mu = \frac{\mu_0}{1 + \frac{\theta}{v_d - v_s}[f(v_g, v_d, v_s) + g(v_b, v_d, v_s)]} \tag{63}$$

where θ is the mobility degradation factor, and f and g can be written as

$$f(v_g, v_d, v_s) = [v_g - V_{FB} - \phi](v_d - v_s) - \frac{1}{2}(v_d^2 - v_s^2) \tag{64}$$

$$g(v_b, v_d, v_s) = \frac{2\gamma}{3}[(v_d - v_b + \phi)^{3/2} - (v_s - v_b + \phi)^{3/2}] \tag{65}$$

where V_{FB} is the flat-band voltage, ϕ is the surface potential, γ is the body-effect coefficient [60]. In the subsequent subsection, the above equations are utilized to analyze the common-mode mode linearization techniques, and to evaluate its effectiveness.

The gate linearization is achieved by applying $v_g = V_G + v_c$ to the gate terminal. As a result, f becomes

$$f(v_{ds}) = [V_G + v_c - V_{FB} - \phi]v_{ds} - \frac{v_{ds}^2}{2} = [V_G - V_{FB} - \phi]v_{ds} \tag{66}$$

Also, g can be expanded by using the Taylor series expansion at $v_{ds} = 0$. This can be achieved by expressing $v_d = v_c + \delta$, $v_s = v_c - \delta$, and $u = v_c - v_b + \phi$ to simplify the

computation. In this case, g can be expressed as

$$g = \frac{2\gamma}{3}u^{3/2}\Big[(1+\frac{\delta}{u})^{3/2} - (1-\frac{\delta}{u})^{3/2}\Big] \tag{67}$$

For $u \gg \delta$, g becomes

$$g = \frac{2\gamma}{3}u^{3/2}\Big[3\frac{\delta}{u} - \frac{\delta^3}{8u^3} - \frac{3\delta^5}{128u^5} + \dots\Big] \tag{68}$$

If the higher order terms are ignored, and for $\delta = v_{ds}/2$, g can be written as

$$g(v_{ds}) = \gamma\Big(v_{ds}(v_c - v_b + \phi)^{1/2} - \frac{v_{ds}^3}{96(v_c - v_b + \phi)^{3/2}}\Big) \tag{69}$$

By using (66) and (69), and the definitions in (9) and (11), the drain current (without the mobility degradation) can be written as

$$I_d = \frac{\mu C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} + \frac{\gamma^4 v_{ds}^3}{96V_{c_1}^3}\right\} \tag{70}$$

To account for the mobility degradation, the mobility can be expressed as

$$\mu = \frac{\mu_0}{1 + \theta(V_{G_1} + V_{c_1} - \frac{\gamma^4 v_{ds}^2}{96V_{c_1}^3})} \tag{71}$$

Using (10) for further simplification, the mobility becomes

$$\mu = \frac{\mu_1}{1 + \theta_1(V_{c_1} - \frac{\gamma^4 v_{ds}^2}{96V_{c_1}^3})} \tag{72}$$

For $\theta_1 \ll 1/(V_{c_1} - \frac{\gamma^4 v_{ds}^2}{96V_{c_1}^3})$, and applying $y = 1/(1+x) \approx 1 - x$ approximation to (72), the drain current for the gate linearization technique can be written as

$$I_d = \frac{\mu_1 C_{ox} W}{L}\Big(v_{ds}(V_G - V_T)(1 - \theta_1 V_{c_1}) + \frac{\gamma^4 v_{ds}^3}{96V_{c_1}^3}(1 + \theta_1(V_{G_1} - 2V_{c_1}))\Big) \tag{73}$$

The common-mode strategy is an extended version of the gate linearization technique. In addition to the common-mode gate signal, the common-mode body signal ($v_b = -V_B + v_c$) is applied to the body terminal. As a result, g in (69) becomes

$$g(v_{ds}) = \gamma[v_{ds}(V_B + \phi)^{1/2} - \frac{v_{ds}^3}{96(V_B + \phi)^{3/2}}] \tag{74}$$

By using (66) and (74), the drain current can be written as

$$I_d = \frac{\mu C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} + \frac{\gamma v_{ds}^3}{(V_B + \phi)^{3/2}}\right\} \tag{75}$$

Moreover, the mobility, in this case, becomes

$$\mu = \frac{\mu_2}{1 - \frac{\theta_2 \gamma v_{ds}^2}{96(V_B+\phi)^{3/2}}} \tag{76}$$

Finally, by applying $\theta_2 \ll \frac{96(V_B+\phi)^{3/2}}{\gamma v_{ds}^3}$ and $y = 1/(1-x) \approx 1 + x$ approximations to (76), the drain current for the common-mode strategy can be approximated as

$$I_d = \frac{\mu_2 C_{ox} W}{L}\left\{[V_G - V_T]v_{ds} + \frac{\gamma(1 + \theta_2[V_G - V_T])}{96\sqrt[3]{V_B + \phi}} v_{ds}^3\right\} \tag{77}$$

APPENDIX B

SPEED ANALYSIS OF BWCDAC AND FGDAC

The speed performance of the BWCDAC and the FGDAC are analyzed and compared by using the illustrated DAC structure in Figure 61. For simplification in the analysis, the output resistances of the voltage-reference and epots are assumed to be much smaller than the on-resistance of the switches, and therefore ignored in the analysis. Also, the time constants of input branches are assumed to be same. Hence, $R_{on_i} = R_{on}/2^{i-1}$ and $C_i = 2^{i-1}C$ for the BWCDAC, and $R_{on_i} = R_{on}$ and $C_i = C$ for the FGDAC, where $i = 1, ..., N$, C is the unit-size capacitance, and R_{on} is the switch on-resistance.

Using the small signal models illustrated in Figure 62, the relation between V_{in}, V_x, and V_{out} can be expressed as

$$\frac{V_{in}C_{eq_1}}{1 + R_{eq_1}C_{eq_1}s} = V_x\Big(C_f + C_{amp} + \frac{C_{eq}}{1 + R_{eq_2}C_{eq_2}s}\Big) - V_{out}C_f \tag{78}$$

where $C_{eq_1} = \Sigma_{i=1}^{N} b_i C_i = k_1 C$, and $R_{eq_1} = R_{on}/k_1$. Similarly, $C_{eq_2} = \Sigma_{i=1}^{N} \overline{b}_i C_i = k_2 C$ and $R_{eq_2} = R_{on}/k_2$. In addition, $C_{eq} = C_{eq_1} + C_{eq_2}$ and $R_{eq_1}C_{eq_1} = R_{eq_2}C_{eq_2} = R_{on}C$. Based on the amplifier type and model used in the analysis, the transfer function of the DACs can be different. In the next subsections, one-stage and two-stage amplifier models are utilized to find the effect of the amplifier on the DAC performances.

B.1 Using one-stage amplifier

A simplified model shown in Figure 62a is used to analyze the DACs with one-stage inverting amplifier. Based on the illustrated model for the amplifier, V_{out} can be expressed in terms of V_x as follows

$$\frac{V_{out}}{V_x}(s) = -\frac{(G_m - sC_f)R_o}{1 + sR_o(C_L + C_f)} \tag{79}$$

where G_m and R_o are the transconductance and output resistance of the amplifier, and C_f and C_L are the feedback and load capacitors. Assuming the amplifier gain, $G_m R_o$, is large

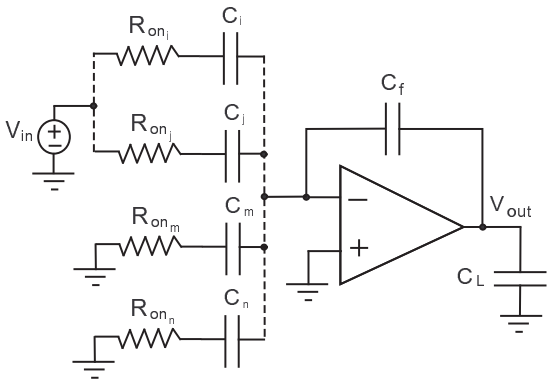


Figure 61. DAC structure used to analyze BWCDAC and FGDAC. C_L and C_f are the load and feedback capacitors, respectively. Also, for ($x = i, j, ..., m, n$), R_{on_x} is the on-resistance of the switches and C_x is the input capacitor. This structure illustrates the connections when some of the input capacitors are connected to the input while others are connected to the ground. For BWCDAC, V_{in} is equal to V_{ref}, while it is equal to V_{epot} for FGDAC. For simplification the output resistances of the reference and epots are assumed to be much less than R_{on}.

enough, the DAC transfer function becomes

$$\frac{V_{out}}{V_{in}}(s) = -\frac{(G_m - sC_f)C_{eq_1}}{G_m C_f + sC_1^2 + s^2 R_{eq_1} C_{eq_1} C_2^2} \tag{80}$$

where C_1^2 and C_2^2 are

$$C_1^2 = C_f(G_m R_{eq_1} C_{eq_1} + C_L) + (C_{amp} + C_{eq})(C_L + C_f) \tag{81}$$

$$C_2^2 = (C_{amp} + C_L)C_f + C_L C_{amp} \tag{82}$$

where C_{amp} is the amplifier input capacitance. Assuming $C_1^4 \gg 4G_m C_f R_{eq_1} C_{eq_1} C_2^2$, the poles of (80) can be approximated as

$$p_1 = -\frac{G_m C_f}{C_1^2} \quad \& \quad p_2 = -\frac{C_1^2}{R_{eq_1} C_{eq_1} C_2^2} \tag{83}$$

Based on this analysis, the time constants, τ_{DAC_1} and τ_{DAC_2}, can be computed as

$$\tau_{DAC_1} = R_{on}C + \frac{C_f C_L + (C_{amp} + C_{eq})(C_f + C_L)}{G_m C_f} \tag{84}$$

$$\tau_{DAC_2} = \frac{R_{on}C(C_f(C_{amp} + C_L) + C_L C_{amp})}{C_f(G_m R_{on} C + C_L) + (C_{amp} + C_{eq})(C_L + C_f)} \tag{85}$$

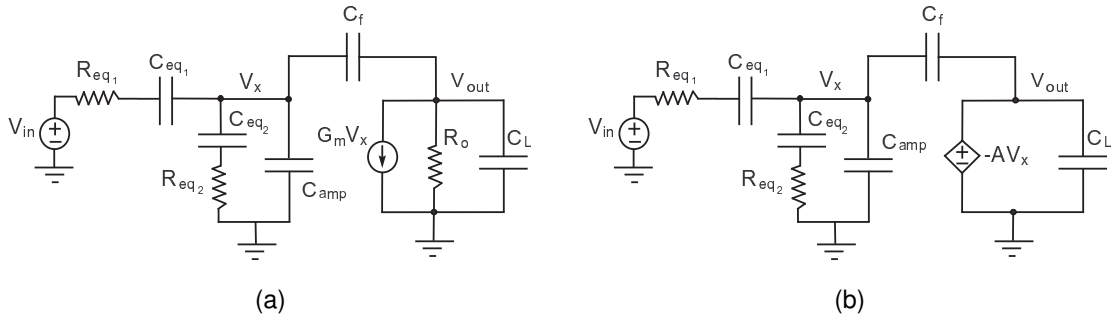


Figure 62. Small signal models used to analyze the DAC structures. R_{eq_1} and C_{eq_1} are the equivalent resistance and capacitance of the selected branches, and R_{eq_2} and C_{eq_2} are the equivalent resistance and capacitance of the unselected branches. Also, C_{amp} is the input capacitance of the amplifier, and C_L is the load capacitor. (a) Using a simplified one-stage amplifier model. G_m and R_o are the transconductance and output resistance of the amplifier. (b) Using a simplified two-stage amplifier model. A is the amplifier gain that has one dominant pole.

For large values of N and small values of G_mR_{on}, the time constants of the BWCDAC, τ_{BWCDAC_1} and τ_{BWCDAC_2}, can be approximated as

$$\tau_{BWCDAC_1} = \frac{C_L + (1 + C_{amp}/(2^N C))(C_L + 2^N C)}{G_m} \tag{86}$$

$$\tau_{BWCDAC_2} = \frac{R_{on}C(2^N C(C_{amp} + C_L) + C_L C_{amp})}{2^N C_L C + (C_{amp} + 2^N C)(C_L + 2^N C)} \tag{87}$$

Similarly, based on these assumptions, the time constants of the FGDAC, τ_{FGDAC_1} and τ_{FGDAC_2}, can be written as

$$\tau_{FGDAC_1} = \frac{C_L + (N + C_{amp}/C)(C_L + C)}{G_m} \tag{88}$$

$$\tau_{FGDAC_2} = \frac{R_{on}C(C(C_{amp} + C_L) + C_L)C_{amp}}{C_L C + (C_{amp} + NC)(C_L + C)} \tag{89}$$

As a result, the speed performances of the BWCDAC and the FGDAC can be compared based on these approximated time constants, and the relationship between τ_{DAC_1} and τ_{DAC_2} is expressed as

$$\tau_{DAC_1} = \frac{1}{\tau_{DAC_2}} \cdot \frac{C_x R_{on} C}{G_m} \tag{90}$$

where $C_x = C_f(C_{amp} + C_L) + C_L C_{amp}$. The above equation implies that the multiplication of the BWCDAC time constants is equal to the multiplication of the FGDAC time constants if the load capacitance, the on-resistance of the switches, the amplifier transconductance, and the unit capacitance of these converters are the same.

B.2 Using two-stage amplifier

When a two-stage amplifier is employed for the DACs, a simplified model illustrated in Figure 62b can be used to express V_{out} in terms of V_x. In this model it is assumed that the second pole and the zero of the amplifier are beyond the gain-bandwidth of the amplifier. Therefore, V_{out} becomes equal to $-A(s)V_x$ for $A(s) = GB/(s + w_a)$, where GB is the gain-bandwidth and w_a is the dominant pole of the amplifier. As a result, for a large amplifier DC gain, the transfer function of the DAC can be written as

$$\frac{V_{out}}{V_{in}}(s) = \frac{-GB \cdot C_{eq_1}}{GB \cdot C_f + sC_1 + s^2 R_{eq_1} C_{eq_1} C_2} \tag{91}$$

where C_1 and C_2 are

$$C_1 = C_f + C_{eq} + C_{amp} + R_{eq_1} C_{eq_1} C_f GB \tag{92}$$

$$C_2 = C_f + C_{amp} \tag{93}$$

The poles of (91) for $C_1^2 \gg 4C_f R_{eq_1} C_{eq_1} C_2 GB$ can be found as

$$p_1 = -\frac{C_f GB}{C_1} \quad \& \quad p_2 = -\frac{C_1}{R_{eq_1} C_{eq_1} C_2} \tag{94}$$

Based on the computed poles, the time constants, τ_{DAC_1} and τ_{DAC_2}, can be expressed as

$$\tau_{DAC_1} = \frac{1}{GB} \cdot \left(1 + R_{on} C \cdot GB + \frac{C_{eq} + C_{amp}}{C_f}\right) \tag{95}$$

$$\tau_{DAC_2} = \frac{R_{on} C}{1 + \frac{C_{eq} + R_{on} C \cdot C_f GB}{C_f + C_{amp}}} \tag{96}$$

The expressions for the time constants of the BWCDAC, τ_{BWCDAC_1} and τ_{BWCDAC_2}, can be simplified for large values of N, and they become

$$\tau_{BWCDAC_1} = \frac{2 + C_{amp}/(2^N C)}{GB} + R_{on}C \tag{97}$$

$$\tau_{BWCDAC_2} = \frac{R_{on}C(2^N C + C_{amp})}{C_{amp} + 2^N C(2 + R_{on}C \cdot GB)} \tag{98}$$

Similarly, the time constants of the FGDAC, τ_{FGDAC_1} and τ_{FGDAC_2}, become

$$\tau_{FGDAC_1} = \frac{N + 1 + C_{amp}/C}{GB} + R_{on}C \tag{99}$$

$$\tau_{FGDAC_2} = \frac{R_{on}C(C + C_{amp})}{C_{amp} + (N + 1)C + R_{on}C^2 \cdot GB} \tag{100}$$

Again, the relationship between τ_{DAC_1} and τ_{DAC_2} becomes

$$\tau_{DAC_1} = \frac{1}{\tau_{DAC_2}} \cdot \frac{R_{on}(C + C_{amp})}{GB} \tag{101}$$

Similar to the one-stage amplifier case, the multiplication of the BWCDAC time constants becomes equal to the multiplication of the FGDAC time constants for the same unit and input amplifier capacitances, on-resistance of switches, and gain-bandwidth of amplifier.

APPENDIX C

NOISE ANALYSIS OF BWCDAC AND FGDAC

To compare the noise performances of the FGDAC and the BWCDAC, one-stage and two stage-amplifier models are utilized in the analysis. The general expression used for the total DAC noise is as follows

$$e_{DAC}^2 = e_{reset}^2 + e_{amp}^2 B_{n_1} A_{n_1} + (e_{ref}^2 + e_{Ron}^2) B_{n_2} A_{n_2} \tag{102}$$

where e_{amp}^2, e_{Ron}^2, and e_{ref}^2 are the broadband noise contribution of the amplifier, the switches, and the reference, and e_{reset}^2 is the kT/C noise introduced during the reset phase of the BWCDAC. B_{n_1} and B_{n_2} are the noise bandwidths of the amplifier and the reference/switches, and A_{n_1} and A_{n_2} are the gain of the DAC from the amplifier and the reference/switches, respectively.

C.1 Using one-stage amplifier

The noise analysis for the one-stage amplifier is done by using the simplified model shown in Figure 63a. The transfer function for V_{out}/V_{in} is given in (80), where C_{eq_1} and R_{eq_1} are now equal to C_{eq} and R_{eq}. This transfer function can be used to express the noise contribution from input to output. Similarly, the noise contribution of the amplifier can be computed by finding the transfer function from V_{amp} to V_{out}. This can be expressed as

$$\begin{aligned} V_x(C_{eq} + C_f + C_{amp} + sR_{eq}C_{eq}(C_f + C_{amp})) + V_{amp}(C_{eq} + C_f + sR_{eq}C_{eq}C_f) \\ = V_{out}C_f(1 + sR_{eq}C_{eq}) \end{aligned} \tag{103}$$

Using the relationship between V_x and V_{out} as given in (79), the transfer function for V_{out}/V_x can be written as

$$\frac{V_{out}}{V_{amp}}(s) = \frac{G_m(C_{eq} + C_f) + sC_a^2 + s^2 R_{eq}C_{eq}C_f C_{amp}}{G_m C_f + sC_{b_1}^2 + s^2 R_{eq}C_{eq}C_{b_2}^2} \tag{104}$$

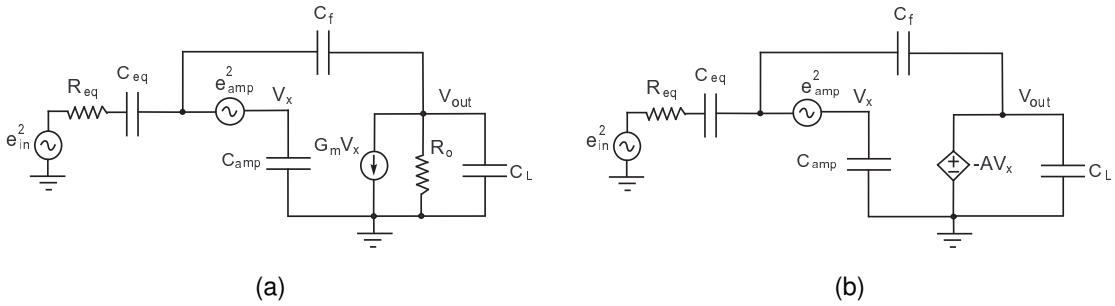


Figure 63. Models used to analyze the noise of the DAC structures. e_{in}^2 and e_{amp}^2 are the input noise and the amplifier noise, respectively. (a) Using a simplified one-stage amplifier model for the DAC amplifier. (b) Using a simplified two-stage amplifier model for the DAC amplifier.

where C_a^2, $C_{b_1}^2$, and $C_{b_2}^2$ are

$$C_a^2 = R_{eq}G_mC_{eq}C_f + C_{amp}C_f \tag{105}$$

$$C_{b_1}^2 = R_{eq}G_mC_{eq}C_f + C_LC_f + (C_{eq} + C_{amp})(C_f + C_L) \tag{106}$$

$$C_{b_2}^2 = C_{amp}(C_f + C_L) + C_LC_f \tag{107}$$

The zeros of this transfer function becomes

$$z_{1,2} = \frac{-C_a^2 \pm C_a^2(1 - \frac{4R_{eq}C_{eq}C_{amp}C_fG_m(C_{eq}+C_f)}{C_a^4})^{0.5}}{2R_{eq}C_{eq}C_{amp}C_f} \tag{108}$$

For $C_a^4 \gg 4R_{eq}C_{eq}C_{amp}C_fG_m(C_{eq} + C_f)$, z_1 and z_2 can be approximated as

$$z_1 \approx -\frac{G_m(C_{eq} + C_f)}{C_a^2} \quad \& \quad z_2 \approx -\frac{C_a^2}{R_{eq}C_{eq}C_{amp}C_f} \tag{109}$$

Similarly, the poles of the transfer function can be found as

$$p_{1,2} = \frac{-C_{b_1}^2 \pm C_{b_1}^2(1 - \frac{4R_{eq}C_{eq}C_fG_mC_{b_2}^2}{C_{b_1}^4})^{0.5}}{2R_{eq}C_{eq}C_{b_2}^2} \tag{110}$$

For $C_{b_1}^4 \gg 4R_{eq}C_{eq}C_fG_mC_{b_2}^2$, p_1 and p_2 becomes

$$p_1 \approx -\frac{G_mC_f}{C_{b_1}^2} \quad \& \quad p_2 \approx -\frac{C_{b_1}^2}{R_{eq}C_{eq}C_{b_2}^2} \tag{111}$$

Depending on the capacitance values, the location of the poles and zeros may change. If we assume that p_1 is the dominant pole of the system, and that $p_1 \ll z_1$, which is true for $G_m R_{on} \ll 1$ and $C_L \gg C_{amp}$. In this case, V_{out}/V_{amp} can be approximated as a single pole system in the bandwidth of interest. Therefore, the transfer function for $|V_{out}/V_{amp}|^2$ can be written as

$$|H_1(jw)|^2 = \Big(\frac{C_{eq} + C_f}{C_f}\Big)^2 \frac{1}{1 + (2\pi f \cdot C_{b_1}^2/(G_m C_f))^2} \tag{112}$$

Based on the above equation, the gain, A_{n_1}, can be expressed as $(C_{eq} + C_f)^2/C_f^2$ and the bandwidth, B_{n_1}, becomes $G_m C_f/(4C_{b_1}^2)$.

Moreover, if the zero of the transfer function in (80), $z = G_m/C_f$, and its second pole, p_2, cancel or are further away than the bandwidth of interest, the transfer function for $|V_{out}/V_{in}|^2$ becomes

$$|H_2(jw)|^2 = \Big(\frac{C_{eq_1}}{C_f}\Big)^2 \frac{1}{1 + (2\pi f \cdot C_1^2/(G_m C_f))^2} \tag{113}$$

As a result, the gain, A_{n_2}, becomes $(C_{eq_1}/C_f)^2$ and the bandwidth, B_{n_2}, can be approximated as $G_m C_f/(4C_1^2)$.

By using these gain and bandwidth expression together with (102), the total thermal noise of the BWCDAC for one-stage amplifier can be expressed as

$$e_{BWC}^2 = \Big(\frac{kT}{C_{eq}}\Big) \cdot \frac{C_{eq}}{C_f} + \Big(4kTR_{eq} + e_{ref}^2 + (\frac{C_{eq} + C_f}{C_{eq}})^2 \cdot e_{amp}^2\Big) \cdot \frac{C_{eq}^2}{C_1^2} \frac{G_m}{4C_f} \tag{114}$$

Using the BWCDAC capacitance values the above expression can be written as

$$e_{BWC}^2 = \frac{kT}{2^N C} + \Big(4kT\frac{R_{on}}{2^N - 1} + e_{ref}^2 + e_{amp}^2 \cdot (\frac{2^{N+1} - 1}{2^N - 1})^2\Big) \cdot \frac{(2^N - 1)^2 G_m C}{2^{N+2} C_1^2} \tag{115}$$

where $C_1^2 \approx 2^N C_L C + (C_{amp} + (2^N - 1)C)(C_L + 2^N C))$ for $G_m R_{eq} C_{eq} = G_m R_{on} C \ll C$. For large values of N, the total noise can be approximated as

$$e_{BWC}^2 = \frac{kT}{2^N C} + \Big(kT\frac{R_{on}}{2^N} + \frac{e_{ref}^2}{4} + e_{amp}^2\Big) \cdot \frac{G_m}{C_x} \tag{116}$$

where $C_x = C_L + (1 + C_{amp}/(2^N C)) \cdot (2^N C + C_L)$.

Similarly, for one-stage amplifier the equivalent output thermal noise of the FGDAC can be approximated as

$$e_{FG}^2 = \left(4kT\frac{R_{on}}{N} + \frac{e_{epot}^2}{N} + e_{amp}^2(\frac{N+1}{N})^2\right) \cdot \frac{N^2C \cdot G_m}{4C_1^2} \tag{117}$$

where e_{epot}^2 is the broadband noise contribution of the epots and $C_1^2 \approx C_LC + (C_{amp} + NC)(C_L + C)$ for $G_mR_{eq}C_{eq} \ll C$. For large values of N, the above expression becomes

$$e_{FG}^2 = \left(4kT\frac{R_{on}}{N} + \frac{e_{epot}^2}{N} + e_{amp}^2\right) \cdot \frac{N^2 \cdot G_m}{4C_y} \tag{118}$$

where $C_y = C_L + (N + C_{amp}/C)(C_L + C)$.

C.2 Using two-stage amplifier

Depending on the amplifier model and its number of stages, the expression for the total output referred noise changes. Therefore, the noise analysis of the DACs with two-stage amplifier is performed by using the simplified model illustrated in Figure 63b. Using this model and (103), it can be shown that

$$\frac{V_{out}}{V_{amp}}(s) = \frac{GB(C_{eq} + C_f + sR_{eq}C_{eq}C_f)}{C_fGB + sC_x + s^2R_{eq}C_{eq}(C_f + C_{amp})} \tag{119}$$

where C_x is

$$C_x = C_fR_{eq}C_{eq}GB + C_{eq} \tag{120}$$

The zero of this transfer function becomes

$$z = -\frac{C_{eq} + C_f}{C_f}\frac{1}{R_{eq}C_{eq}} \tag{121}$$

$$p_{1,2} = \frac{-C_x \pm C_x(1 - \frac{4R_{eq}C_{eq}C_fGB(C_f+C_{amp})}{C_x^2})^{0.5}}{2R_{eq}C_{eq}(C_f + C_{amp})} \tag{122}$$

and for $C_x^2 \gg 4C_{eq}R_{eq}C_fGB(C_f + C_{amp})$, the poles can approximated as

$$p_1 \approx -\frac{C_f}{C_x^2}GB \quad \& \quad p_2 \approx -\frac{C_f}{C_f + C_{amp}}GB \tag{123}$$

Assuming $GB < 1/(R_{eq}C_{eq})$ and $z \approx p_1$, $|V_{out}/V_{amp}|^2$ becomes

$$|H_1(jw)|^2 = \frac{(C_{eq} + C_f)^2/C_f^2}{1 + \left(2\pi f \cdot (C_f + C_{amp})/(C_f GB)\right)^2} \tag{124}$$

The above equation yields $A_{n_1} = ((C_{eq} + C_f)/C_f)^2$ and $B_{n_1} = C_f GB/(4(C_f + C_{amp}))$.

Also, the effect of the input noise can be used found by using (91) and assuming that the first pole of (91) is the dominant pole in the bandwidth of interest. Then this transfer function can be approximated as a single pole system, and thus $|V_{out}/V_{in}|^2$ can be written as

$$|H_2(jw)|^2 = \frac{(C_{eq_1}/C_f)^2}{1 + \left(2\pi f \cdot C_1/(C_f GB)\right)^2} \tag{125}$$

Therefore, the gain and the bandwidth become $A_{n_2} = (C_{eq_1}/C_f)^2$ and $B_{n_2} = C_f GB/(4C_1)$. As a result, the total thermal noise of the BWCDAC for two-stage amplifier becomes

$$e_{BWC}^2 = \frac{kT}{C_{eq}}\frac{C_{eq}}{C_f} + (4kTR_{eq} + e_{ref}^2) \cdot \frac{C_{eq}^2 GB}{4C_1 C_f} + e_{amp}^2 \cdot \frac{(C_{eq} + C_f)^2 GB}{4C_f(C_f + C_{amp})} \tag{126}$$

Using the capacitance values of the BWCDAC, the above expression can be rewritten as

$$e_{BWC}^2 = \frac{kT}{2^N C} + (4kT\frac{R_{on}}{2^N - 1} + e_{ref}^2) \cdot \frac{(2^N - 1)^2 C \cdot GB}{2^{N+2} C_1} + e_{amp}^2 \cdot \frac{(2^{N+1} - 1)^2 C \cdot GB}{2^{N+2}(2^N C + C_{amp})} \tag{127}$$

where $C_1 \approx (2^{N+1} - 1)C + C_{amp}$. For large values of N, the total thermal noise of the BWCDAC becomes

$$e_{BWC}^2 = \frac{kT}{2^N C} + (4kT\frac{R_{on}}{2^N} + e_{ref}^2) \cdot \frac{2^{N-2} C \cdot GB}{(2^{N+1} C + C_{amp})} + e_{amp}^2 \cdot \frac{2^N C \cdot GB}{(2^N C + C_{amp})} \tag{128}$$

Lastly, for two stage amplifier, the equivalent thermal noise of the FGDAC becomes

$$e_{FG}^2 = (4kTR_{on}N + e_{epot}^2) \cdot \frac{NC \cdot GB}{4((N + 1)C + C_{amp})} + e_{amp}^2 \cdot \frac{(N + 1)^2 C \cdot GB}{4(C + C_{amp})} \tag{129}$$

REFERENCES

[1] A. B. R. M. Edenfeld, D. Kahng and Y. Zorian, "2003 technology roadmap for semiconductors," *IEEE Computer*, vol. 37, 2004.

[2] C. Teuscher, I. Sheng, S.and ODonnell, K. Stone, and R. Brodersen, "Design and implementation issues for a wideband indoor DS-CDMA system providing multimedia access," *Proc. 34th Annu. Allerton Conf. Communication, Control, and Computing*, 1996.

[3] N. Zhang, C. Teuscher, H. Lee, and B. Brodersen, "Architectural implementation issues in a wideband receiver using multiuser detection," *Proc. 36th Annu. Allerton Conf. Communication, Control, and Computing*, 1998.

[4] M. Caudill and C. Butler, *Understanding Neural Networks: Volume 1: Basic Networks*. Cambridge, Massachusetts: The MIT Press, 1992.

[5] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 7, 1943.

[6] S. K. Singh, R. W. Newcomb, P. Gomez, and V. Rodellar, "A means of VLSI current controlled weight setting in ANNs," *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, 1995.

[7] S. Sakurai and M. Ismail, "A CMOS square-law programmable floating resistor independent of the threshold voltage," *IEEE Trans. on Circuit and Systems II: Analog and Digital Signal Processing*, vol. 39, 1992.

[8] S. P. Singh, J. V. Hanson, and J. Vlach, "A new floating resistor for CMOS technology," *IEEE Trans. on Circuit and Systems*, vol. 36, 1989.

[9] Y. Tsividis, M. Banu, and J. Khoury, "Continuous-time MOSFET-C filters in VLSI," *IEEE Journal of Solid State Circuits*, vol. SC-21, Feb. 1986.

[10] M. Kushima, H. Tanno, K. Kumagai, and O. Ishizuka, "Low-power and wide-input range voltage controlled linear variable resistor using an FG-MOSFET and its application," *The International Conference on Circuits/Systems Computers and Communications*, July 2002.

[11] H. Youssef, R. Newcomb, and M. Zaghloul, "A CMOS voltage-controlled linear resistor with wide dynamic range," *Twenty-First Southeastern Symposium on System Theory*, 1989.

[12] Z. Wang and W. Guggenbuhl, "A voltage controllable linear MOS transconductor using bias offset technique," *IEEE Journal of Solid State Circuits*, vol. 25, 1990.

[13] M. Banu and Y. Tsividis, "Fully integrated active RC filters in MOS technology," *IEEE International Solid-State Circuits Conference*, vol. 26, 1983.

[14] Z. Czarnul and Y. Tsividis, "Implementation of MOSFET-C filters based on active RC prototypes," *Electronic Letters*, 1988.

[15] J. N. Babanezhad and G. C. Temes, "A linear NMOS depletion resistor and its application in an integrated amplifier," *IEEE Journal of Solid-State Circuits*, vol. 19, 1984.

[16] Z. Czarnul, "A linear NMOS depletion resistor and its application in an integrated amplifier," *IEEE Journal of Solid-State Circuits*, vol. 22, 1987.

[17] K. Nay and A. Budak, "A voltage-controlled resistance with wide dynamic range and low distortion," *IEEE Trans. on Circuit and Systems*, vol. 30, 1983.

[18] G. Wilson and P. K. Chan, "Analysis of nonlinearities in MOS floating resistor networks," *IEE Proc.-Circuits Devices Syst*, vol. 141, 1994.

[19] L. Sellami, S. Singh, R. Newcomb, A. Rasmussen, and M. Zaghloul, "CMOS bilateral linear floating resistors for neural-type cell arrays," *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers*, vol. 2, 1997.

[20] A. Rasmussen and M. Zaghloul, "CMOS analog implementation of cellular neural network to solve partial differential equations with a microelectromechanical thermal interface," *Proceedings of the 40th Midwest Symposium on Circuits and Systems*, vol. 2, 1997.

[21] S. Tantry, T. Yoneyama, and H. Asai, "Two floating resistor circuits and their applications to synaptic weights in analog neural networks," *IEEE International Symposium on Circuits and Systems*, vol. 1, May 2001.

[22] Z. Czarnul and S. Tagaki, "Design of linear tunable CMOS differential transconductor cells," *Electronics Letters*, vol. 26, 1990.

[23] A. Nedungadi and T. Viswanathan, "Design of linear CMOS transconductance elements," *IEEE Transactions on Circuit and Systems*, 1984.

[24] G. Han and E. Sanchez-Sinencio, "CMOS transconductance multipliers: a tutorial," *Analog and Digital Signal Processing, IEEE Transactions on Circuits and Systems II*, vol. 45, 1998.

[25] P. Allen and D. Holberg, *CMOS Analog Circuit Design*. Oxford University Press, Oxford, 2002.

[26] J. McCreary and P. Gray, "All-MOS charge redistribtion Analog-to-Digital conversion tecniques-part 1," *IEEE Journal of Solid-State Circuits*, vol. SC-10, December 1975.

[27] Y. Yee, L. Terman, and L. Heller, "A two-stage weighted capcitor network for D/A-A/D conversion," *IEEE Journal of Solid-State Circuits*, vol. SC-14, August 1979.

[28] S. Singh, A. Prabhaker, and A. Bhattacharyya, "C-2C ladder-based D/A converters for PCM codecs," vol. 22, pp. 1197–1200, December 1987.

[29] J. McCreary, "Matching properties, and voltage and temperature dependence of MOS capacitors," vol. SC-16, pp. 608–616, December 1981.

[30] J.-B. Shyu, G. Temes, and F. Krummenacher, "Random error effects in matched MOS capacitors and current sources," vol. 19, pp. 948–956, December 1984.

[31] J.-B. Shyu, G. Temes, and K. Yao, "Random errors in MOS capacitors," vol. 17, pp. 1070–1076, December 1982.

[32] S. Singh and A. Bhattacharyya, "Matching properties of linear MOS capacitors," vol. 36, pp. 465–467, March 1989.

[33] A. Hastings, *The art of analog layout*. Prentice-Hall Inc., 2000.

[34] B. Leung and S. Sutarja, "Multi-bit sigma-delta A/D converter incorporating a novel class of dynamic element matching techniques," vol. 39, pp. 35–51, January 1992.

[35] R. Baird and T. Fiez, "Improved ΔΣ DAC linearity using data weighted averaging," vol. 1, pp. 13–16, May 1995.

[36] R. Schreier and B. Zhang, "Noise-shaped multibit D/A convertor employing unit elements," vol. 31, pp. 1712–1713, September 1995.

[37] I. Galton, "Noise-shaping D/A converters for ΔΣ modulation," vol. 1, pp. 441–444, May 1996.

[38] J. Goes, J. Vital, and J. Franca, "Systematic design for optimization of high-speed self-calibrated pipelined A/D converters," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 45, 1998.

[39] T. Brooks, D. Robertson, D. Kelly, A. Del Muro, and S. Harston, "A cascaded sigma-delta pipeline A/D converter with 1.25 MHz signal bandwidth and 89 dB SNR," *IEEE Journal of Solid-State Circuits*, vol. 32, 1997.

[40] J. Candy, "A use of double integration in sigma delta modulation," *IEEE Transactions on Communications*, March 1988.

[41] B. Rothenberg, S. Lewis, and P. Hurst, "A 20-Msample/s switched-capacitor finite-impulse-response filter using a transposed structure," vol. 30, pp. 1350–1356, 1995.

[42] P. Wong and R. Gray, "FIR filters with sigma-delta modulation encoding," vol. 38, pp. 979–990, 1990.

[43] G. Fischer, "Analog FIR filters by switched-capacitor techniques," vol. 37, pp. 808–814, 1990.

[44] Y. L. Cheung and A. Buchwald, "A sampled-data switched-current analog 16-tap fir filter with digitally programmable coefficients in 0.8m cmos," pp. 54–55, 1997.

[45] S. Berg, P. Hurst, S. Lewis, and P. Wong, "A switched-capacitor filter in 2 μm CMOS using parallelism to sample at 80 MHz," pp. 62–63, 1994.

[46] Y. Lee and K. Martin, "A switched-capacitor realization of multiple FIR filters on a single chip," pp. 536–542, 1988.

[47] Q. Huang and G. Moschytz, "Analog FIR filters with an oversampled σ-$\triangle$ modulator," vol. 39, no. 9, pp. 658–663, 1992.

[48] Q. Huang, G. Moschytz, and T. Burger, "A 100 tap FIR/IIR analog linear-phase low-pass filter," pp. 91–92, 1995.

[49] N. Benvenuto, L. Franks, and J. Hill, F., "Realization of finite impulse response filters using coefficients +1, 0, and -1," vol. 33, pp. 1117–1125, 1985.

[50] S. Abeysekera and K. Padhi, "Design of multiplier free FIR filters using a LADF sigma-delta (σ-$\triangle$) modulator," vol. 2, pp. 65–68, 2000.

[51] K. Bult and G. Geelen, "An inherently linear and compact MOST-only current division technique," vol. 27, pp. 1730–1735, 1992.

[52] A. Chiang, "Low-power adaptive filter," pp. 90–91, 1994.

[53] W. Figueroa, D. Hsu, and C. Diorio, "A mixed-signal approach to high-performance low-power linear filters," vol. 36, pp. 816–822, 2001.

[54] S. Lyle, G. Worstell, and R. Spencer, "An analog discrete-time transversal filter in 2.0 μm CMOS," vol. 2, pp. 970–974, 1992.

[55] X. Wang and R. Spencer, "A low-power 170-MHz discrete-time analog FIR filter," vol. 33, pp. 417–426, 1998.

[56] F. Farag, C. Galup-Montoro, and M. Schneider, "Digitally programmable switched-current FIR filter for low-voltage applications," vol. 35, pp. 637–641, 2000.

[57] P. Sirisuk, A. Worapishet, S. Chanyavilas, and K. Dejhan, "Implementation of switched-current FIR filter using distributed arithmetic technique: exploitation of digital concept in analogue domain," vol. 1, pp. 143–148, 2004.

[58] W. R. Patterson and F. S. Shoucair, "Harmonic suppression in unbalanced analogue MOSFET circuit topologies using body signals," *Electronic Letters*, vol. 25, Dec. 1989.

[59] H. K. J. Ihantola and J. L. Moll, "Design theory of a surface field effect transistors," *Solid-State Electron.*, vol. 7, 1964.

[60] M. H. White, F. Van De Wiele, and J. P. Lambot, "High accuracy MOS models for computer-aided design," *IEEE Trans.*, vol. ED-27, 1980.

[61] Y. P. Tsividis, *Operation and modelling of the MOS transistor*. New York: McGraw-Hill Companies, Inc., 1987.

[62] C. Bleiker and H. Melchior, "A four-state eeprom using floating-gate memory cell," *IEEE J. Solid State Circuits*, vol. 22, no. 3, 1987.

[63] H. Nozama and S. Kokyama, "A thermionic electron emission model for charge retention in SAMOS structures," *Japanese Journal of Applied Physics*, vol. 21, 1992.

[64] T. Shibata and T. Ohmi, "A functional MOS transistor featuring gatelevel weighted sum and threshold operations," *IEEE Trans. Electron Devices*, 1992.

[65] T. S. Lande, E. Olsen, and C. Toumazou, "Resistive equivalents in CMOS," *Electronic Letters*, vol. 39, 2003.

[66] E. Özalevli and P. Hasler, "Design of a CMOS floating-gate resistor for highly linear amplifier and multiplier applications," *Proceedings of Custom Integrated Circuits Conference, San Jose, California*, 2005.

[67] E. Özalevli and P. Hasler, "Programmable floating-gate CMOS Resistors," *Proceedings of International Symposium on Circuits and Systems, Kobe, Japan*, 2005.

[68] K. Vavelidis and Y. Tsividis, "R-MOSFET structure based on current division," *Electronics Letters*, vol. 29, 1993.

[69] K. Vavelidis, Y. P. Tsividis, F. O. Eynde, and Y. Papananos, "Six terminal MOSFET's: Modelling and applications in highly linear, electronically tunable resistors," *IEEE Journal of Solid State Circuits*, vol. 32, Jan. 1997.

[70] Y. Sugimoto, "A 1 V operational, 20 MS/s and 57 db of S/N, current-mode CMOS sample-and-hold IC," *Symposium on VLSI Circuits*, 2001.

[71] K. Hadidi and A. Khoei, "A highly-linear cascode-driver CMOS source follower buffer," *Proceedings of IEEE International Conference on Electronics, Circuits, and Systems*, 1996.

[72] S. Willingham, K. Martin, and A. Ganesan, "A BiCMOS low-distortion 8-MHz low-pass filter," *IEEE Journal of Solid-State Circuits*, vol. 28, 1993.

[73] R. Harrison, J. Bragg, P. Hasler, B. Minch, and S. Deweerth, "A CMOS Programmable Analog Memory Cell Array using Floating-Gate Circuits," *IEEE Trans. on Circuit and Systems*, 2001.

[74] E. Özalevli, P. Hasler, and F. Adil, "Programmable voltage-output, floating-gate Digital-to-Analog converter," in *IEEE Symposium on Circuits and Systems*, (Vancouver, OK), May 2004.

[75] E. Özalevli and P. Hasler, "10-bit programmable voltage-output Digital-Analog converter," *Proceedings of International Symposium on Circuits and Systems, Kobe, Japan*, 2005.

[76] W. Black, D. Allstot, and R. Reed, "A high performance low power CMOS channel filter," vol. 15, pp. 929–938, 1980.

[77] K. Brehmer and J. Wieser, "Large swing CMOS power amplifier," vol. SC-18, pp. 624–629, 1983.

[78] S. Merchant and B. Rao, "Distributed arithmetic architecture for image coding," pp. 74–77, 1989.

[79] H. Cao and W. Li, "VLSI implementation of vector quantization using distributed arithmetic," vol. 2, pp. 668–671, 1996.

[80] M. Sun, T. Chen, and A. Gotlieb, "VLSI implementation of a 16x16 discrete cosine transform," vol. CAS-36, pp. 610–617, 1989.

[81] D. J. Allred, H. Yoo, V. Krishnan, W. Huang, and D. V. Anderson, "Lms adaptive filters uisng distributed arithmetic for high throughput," *IEEE Trans. on Circuit and Systems*, pp. 1327–1337, July 2005.

[82] A. Croisier, D. Esteban, M. Levilion, and V. Rizo, "Digital filter for PCM encoded signals," 1973.

[83] A. Peled and B. Liu, "A new hardware realization of digital filters," vol. 22, pp. 456–462, 1974.

[84] S. A. White, "Applications of distributed arithmetic to digital signal processing: A tutorial review," vol. 6, pp. 4–19, 1989.

[85] P. Lim and B. Wooley, "A high-speed sample-and-hold technique using a miller hold capacitance," vol. 26, pp. 643–651, 1991.

[86] K. Bult and G. Geelen, "A fast-settling CMOS opamp for SC circuits with 90-db DC-gain," vol. 25, pp. 1379–1384, 1990.