HIGH PERFORMANCE RF AND BASEBAND ANALOG-TO-DIGITAL INTERFACE FOR MULTI-STANDARD/WIDEBAND APPLICATIONS

A Dissertation

by

HENG ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Electrical Engineering

HIGH PERFORMANCE RF AND BASEBAND ANALOG-TO-DIGITAL INTERFACE FOR MULTI-STANDARD/WIDEBAND APPLICATIONS

A Dissertation

by

HENG ZHANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Edgar Sánchez-Sinencio
Committee Members,	José Silva-Martínez
	Peng Li
	Alexander Parlos
Head of Department,	Costas N. Georghiades

December 2010

Major Subject: Electrical Engineering

ABSTRACT

High Performance RF and Basdband Analog-to-Digital Interface for
Multi-standard/Wideband Applications. (December 2010)

Heng Zhang, B.S., Peking University, Beijing, China
Chair of Advisory Committee: Dr. Edgar Sánchez-Sinencio

The prevalence of wireless standards and the introduction of dynamic standards/applications, such as software-defined radio, necessitate the next generation wireless devices that integrate multiple standards in a single chip-set to support a variety of services. To reduce the cost and area of such multi-standard handheld devices, reconfigurability is desirable, and the hardware should be shared/reused as much as possible. This research proposes several novel circuit topologies that can meet various specifications with minimum cost, which are suited for multi-standard applications. This doctoral study has two separate contributions: 1. The low noise amplifier (LNA) for the RF front-end; and 2. The analog-to-digital converter (ADC).

The first part of this dissertation focuses on LNA noise reduction and linearization techniques where two novel LNAs are designed, taped out, and measured. The first LNA, implemented in TSMC (Taiwan Semiconductor Manufacturing Company) 0.35µm CMOS (Complementary metal-oxide-semiconductor) process, strategically combined an inductor connected at the gate of the cascode transistor and the capacitive cross-coupling to reduce the noise and nonlinearity contributions of the cascode transistors. The proposed technique reduces LNA NF by 0.35 dB at 2.2 GHz and increases its IIP3 and voltage gain by 2.35 dBm and 2dB respectively, without a compromise on power consumption. The second LNA, implemented in UMC (United Microelectronics Corporation) 0.13μm CMOS process, features a practical linearization technique for high-frequency wideband applications using an active nonlinear resistor, which obtains a robust linearity improvement over process and temperature variations. The proposed linearization method is experimentally demonstrated to improve the IIP3 by 3.5 to 9 dB over a 2.5–10 GHz frequency range. A comparison of measurement results with the prior published state-of-art Ultra-Wideband (UWB) LNAs shows that the proposed linearized UWB LNA achieves excellent linearity with much less power than previously published works.

The second part of this dissertation developed a reconfigurable ADC for multi-standard receiver and video processors. Typical ADCs are power optimized for only one operating speed, while a reconfigurable ADC can scale its power at different speeds, enabling minimal power consumption over a broad range of sampling rates. A novel ADC architecture is proposed for programming the sampling rate with constant biasing current and single clock. The ADC was designed and fabricated using UMC 90nm CMOS process and featured good power scalability and simplified system design. The programmable speed range covers all the video formats and most of the wireless communication standards, while achieving comparable Figure-of-Merit with customized ADCs at each performance node. Since bias current is kept constant, the reconfigurable ADC is more robust and reliable than the previous published works.

DEDICATION

To my parents

ACKNOWLEDGEMENTS

Doctoral work is usually considered as independent research, but it would not have been done without the help from many others. Here I wish to acknowledge these inspiring and gracious people I have met during my graduate study years.

First of all, I would like to express my sincere appreciation to my advisor, Dr. Edgar Sánchez-Sinencio, who has encouraged and guided me through all my research work. By sharing his vision in this field and showing the serious scientific attitude, he taught me how to be a scholar, an engineer, and a teacher. He has broadened my horizon, fostered my learning skills, and most importantly, helped me to develop my future career goal. I will always remember his precious advice and have him as my role model.

I am grateful to Dr. José Silva-Martínez. He showed me the art of analog design through his in-depth knowledge and pioneering expertise. I learned a lot from his courses and his valuable suggestions to my research.

My sincere gratitude goes to Dr. Qunying Li, my internship mentor at Texas Instruments. He is the one who first guided me to the wonderland of ADCs and taught me how to enjoy the beauty/art of analog IC design. He showed me the true spirit of a Ph.D. scholar, which motivated me throughout my doctoral studies. I would like to thank Dr. Louis Luh, my internship mentor at United Microelectronics Corporation (UMC). He gave me lots of suggestions on how to build a career path, which will definitely help me to succeed in my future career. I would like to acknowledge Dr. KC Wang, my

manager in UMC, for his kind help on the chip fabrications. I thank Ann, my colleague in UMC, for her precious advice on the layout.

I would like to thank my colleague, Xiaohua Fan, for his guidance on the LNA projects, and the discussions which inspired my enthusiasm. His dedication with research and courage towards challenges has stimulated me to work hard towards the goal instead of slacking off. I would not have gone this far in my Ph.D. study without his contributions and help. I would like to thank my team members on the ADC project, including Chao Zhang, Junhua Tan, and Hongbo Chen, for their cooperation that helped me finish my work and made the research a joyful and rewarding learning experience. The hard times and laughter we shared together are truly memorable.

I thank my colleagues at the Analog & Mixed Signal Center for fruitful discussions and sharing ideas, especially Weiji Ho, Lei Chen, Rida Assaad, Erik Pankratz, Marvin Onagajo, Amit Gupta, Hesam Amir-Aslanzadeh, Sang Wook Park, and Cho-Ying Lu. I feel privileged to have worked with these genuine and talented friends. I am grateful to my seniors Haitao Tong, Shanfeng Cheng, Jianhong Xiao, Bo Xia, Jinghua Li, and Chunyu Xin, for the wise advices and continuous support in these years.

My thanks go to Dr. Peng Li and Dr. Alexander Parlos for serving on my committee. Thanks for the kind help from Ms. Ella Gallagher.

Finally, I would like to thank my boyfriend, Sichao Chen, for his endurance, kindness, and encouragement. I’d like to thank my parents for shaping me into who I am today. What they have taught me has been, and will always be, invaluable lessons throughout my life. My love and gratitude for them is beyond words and will last forever.

TABLE OF CONTENTS

Page

CHAPTER Page

Page

LIST OF FIGURES

Page

Page

Page

Page

Page

Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Motivation

The emerging 4G wireless communication system introduces an increasing demand of new services, hence a prevalence of wireless standards, to provide more functionality for the end users. As shown in Fig. 1.1, the next generation wireless devices are likely to support these features simultaneously:

1) Cell phone segment: GSM (global system for mobile communications), GPRS (general packet radio service), UMTS (universal mobile telecommunication system), DECT (digital European cordless telephone), EDGE (enhanced data rate for GSM evolution), AMPS (advanced mobile phone systems), IS-95 (digital version of AMPS), etc.
2) Wireless connectivity segment: Bluetooth (i.e. IEEE 802.15) and Zigbee(i.e. 802.15.4) for the personal area network (PAN), WiFi(i.e. IEEE 802.11a/b/g/n) for the local area network (LAN), UWB(i.e.IEEE 802.15.3a) for short range high data rate applications, WiMAX(i.e. IEEE 802.16) and IEEE 802.20 for the metropolitan area network (MAN).

This dissertation follows the style of *IEEE Journal of Solid State Circuits*.

3) Satellite communication segment: GPS (global system for mobile communication) for navigation.

4) Entertainment segment: FM/XM radio and DVB-T/H.

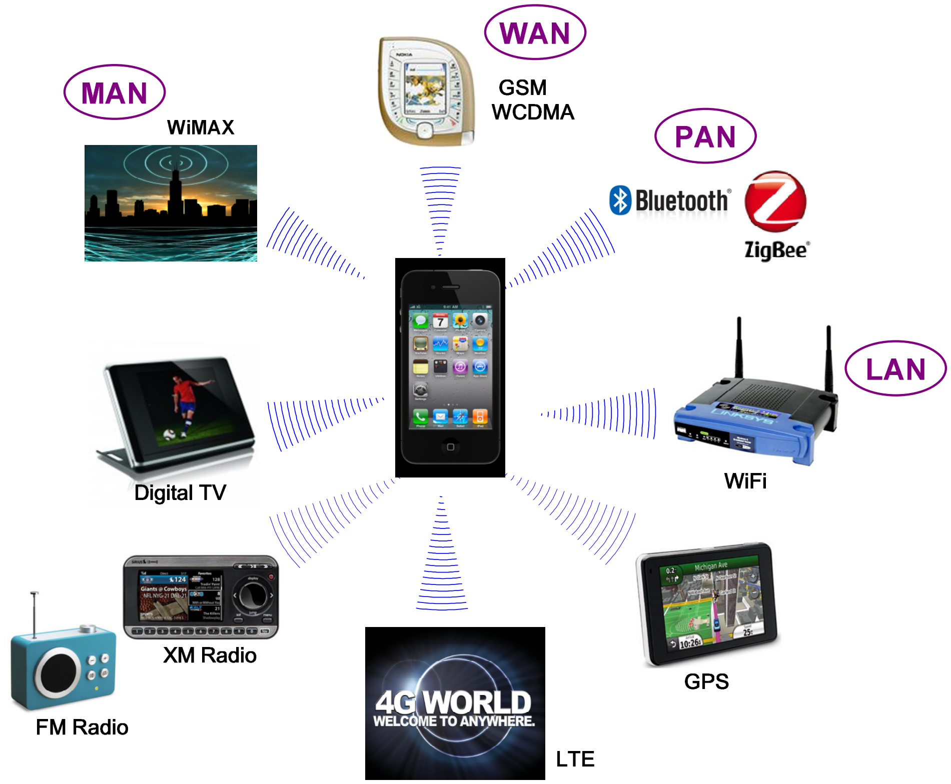


Fig. 1.1. Next generation wireless device.

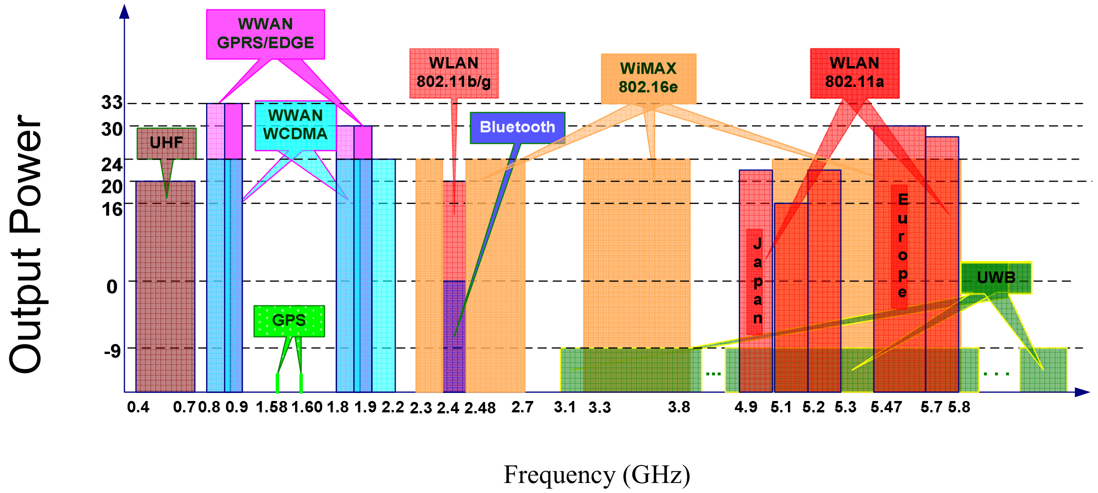


Fig. 1.2. Frequency spectrum of difference services
(Plots courtesy of Camille Chen@Intel).

To reduce power, area, and increase the competitiveness of this new device, it is desired to integrate multiple standards into a single chip-set. Fig. 1.2 shows the frequency spectrum for multiple standards. Two observations can be made: 1) each standard has different definitions of signal power and frequency bands; 2) Hundreds of channels could enter the receiver without any pre-filtering, acting as in-band interferences and creating severe distortion.

Two key challenges can be identified based on these observations: 1) the building blocks in a multi-standard device should satisfy different specifications (e.g. signal-to-noise ratio (SNR) and bandwidth). 2) High linearity must be maintained over a wide frequency range, lest the signal-to-noise/distortion ratio (SNDR) would be dominated by distortion instead of noise.

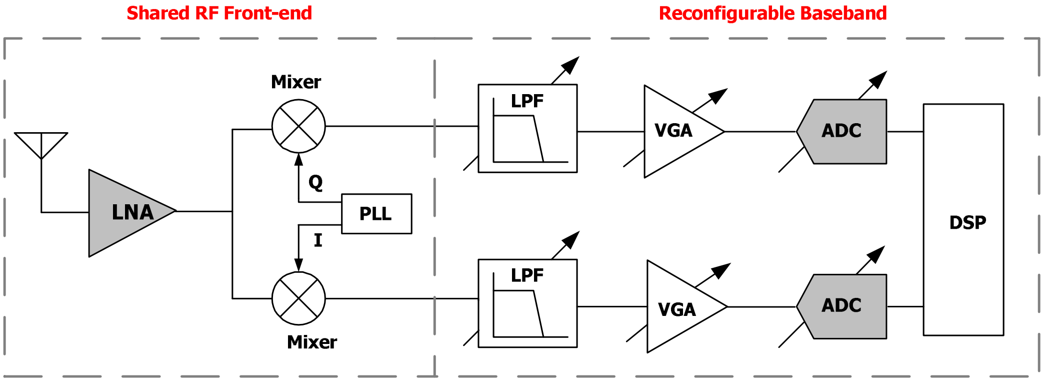


Fig. 1.3. Block diagram of a reconfigurable direct conversion receiver.

How to obtain this goal? A straight approach is to employ parallel narrowband receiver paths with band selection switches, but this increases the cost, area, and power. To optimize the silicon area and power consumption, a highly linear broadband RF front-end with reconfigurable baseband blocks that can meet various specifications with minimum hardware implementation, is a more versatile and cost effective solution [1].

Fig. 1.3 illustrates a reconfigurable direct conversion wireless receiver architecture. The wideband low noise amplifier (LNA) is the first block in the receiver front-end; it amplifies the incoming signal with minimum noise added while providing sufficient dynamic range. The wideband mixer down converts the signal to baseband. The wide tuning range phase locked loop (PLL) selects the signal channel. The low pass filter (LPF) removes the unwanted frequency components form the signal. The variable gain amplifier (VGA) adjusts the signal power to the proper level. The ADC is the last block in the receiver; it is the bridge to connect the analog frontend and the baseband digital signal processing (DSP) module. By digitizing the analog signals with sufficient resolution and speed, the ADC ensures the signal to be processed robustly and reliably in the digital domain.

1.2 Research Contribution

This research investigates two challenging building blocks in the multi-standard receiver, the LNA and the ADC (gray colored in Fig. 1.3), with focuses on linearization techniques, ultra-wideband methods, and power-efficient reconfiguration methodologies. The main goals of this work are summarized as following:

1) Catalogues and analyzes previously reported CMOS LNA linearization techniques. Addresses broadband-LNA-linearization issues for reconfigurable multi-standard/wideband transceivers. Highlights the impact of CMOS technology scaling on linearity and outlines how to design a linear LNA in a deep submicron process. Provides a general design guideline for high-linearity LNAs.
2) Proposes a linearization and noise reduction technique for a differential cascode LNA.

3) Proposes a practical linearization technique for a UWB LNA.

4) Explores the driving forces and new trends for next generation ADCs.

5) Proposes a global offset cancellation technique for a low power cyclic ADC.

6) Proposes a speed reconfigurable, power scalable ADC.

1.3 Dissertation Organization

Chapter II compiles a thorough tutorial of LNA linearization techniques, intuitively explains their operations, and addresses the emerging issues in new applications and advanced technologies.

Chapter III presents the proposed linearization technique for a differential cascode LNA, theoretically analyzes its enhanced performance, with experimental verification through a test chip fabricated in TSMC 0.35μm CMOS process.

Chapter IV describes the proposed linearization technique for a UWB LNA, analyzes its performance using Volterra series, and demonstrates the effectiveness through three UWB LNAs fabricated in UMC 0.13μm CMOS process.

Chapter V explores the next generation ADCs, indentifies new applications and technology scaling as the two main driving forces, and projects the adaptive ADCs, ultra-low power ADCs, and time-domain ADCs as the three new trends. A low power/small area implementation of a cyclic ADC is proposed and verified through a test chip in TI 0.35μm CMOS process.

Chapter VI discusses the proposed speed reconfigurable power scalable ADC, covers the system design and circuit implementations. Chapter VII reveals the layout considerations and lab measurement results for the ADC chip implemented in UMC 90nm CMOS process.

Chapter VIII summarizes this research and discusses the future work.

CHAPTER II

LINEARIZATION TECHNIQUES FOR CMOS LOW NOISE AMPLIFIERS: A TUTORIAL *

2.1 Introduction

The plethora of wireless-communication standards employed in a single geographic region and moreover occupying narrow frequency bands tightly constrains RF-system linearity. Furthermore, the trend in radio research is to simplify/eliminate the expensive front-end module (FEM), which demands a highly linear receiver. In particular, since the low noise amplifier (LNA) is the first block in the receiver chain, it must be sufficiently linear to suppress interference and maintain high sensitivity.

LNA linearization methods should be simple, consume minimum power, and should preserve a low noise figure (NF), gain, and input matching. Many traditional linearization techniques used in lower frequencies are not feasible for LNAs. For example, resistive source degeneration and floating-gate input attenuation reduce the gain and worsen NF or input matching. Hence, LNA linearization proves significantly more challenging than that of baseband circuits [2], often requiring innovative techniques.

Growing research on reconfigurable multi-band/multi-standard and broadband

*©[2010] IEEE. Reprinted, with permission, from "Linearization Techniques for CMOS Low Noise Amplifiers: A Tutorial", by Heng Zhang, and Edgar Sánchez-Sinencio, *IEEE Transactions on Circuits and Systems, Part I: Regular Papers*, December 2010.

transceivers such as ultra-wideband (UWB) and digital TV tuners has propelled interest in broadband LNA design. Radios in the same platform interfere with each other, and multiple channels applied simultaneously to an LNA without filtering act as in-band interferences. Consequently, broadband LNAs must maintain sufficient linearity over a wide frequency range. Emphasis on highly linear transceivers has sparked recent interest in linearizing LNAs [3]. Even though most previously reported techniques target narrowband applications and principally improve only the third-order intercept point (IIP3), we demonstrate why broadband systems require high second-order intercept point (IIP2) and 1dB compression point (P_{1dB}) as well. Because a broadband LNA is exposed to a wide frequency range, we investigate the dependence of IIP2/IIP3 on two-tone (center) frequency and frequency spacing.

Since LNAs typically have low-amplitude, high frequency inputs, the amplifier operates as a weakly nonlinear system with few relevant harmonics (typically only 2nd and 3rd). Thus, Volterra-series analysis [4] can capture the frequency-dependent distortion of LNAs and provide insight into how to compensate that distortion.

CMOS is the most promising technology for systems on a chip. Although MOSFETs are intrinsically more linear than bipolar transistors, they require higher DC current to achieve the necessary transconductance and linearity, thus linearization techniques must be employed to reduce the DC power. Deep-submicron (DSM) technology challenges include nonlinear output conductance, mobility degradation, velocity saturation, and poly-gate depletion; which complicate CMOS LNA linearization, especially in the face of low supply voltages. We present multidimensional

Taylor analysis to evaluate the effects of these nonidealities.

This chapter is organized as follows. Section 2.2 analyzes previously reported CMOS LNA linearization techniques. Section 2.3 discusses new broadband-LNA-linearization issues arising in multi-band/multi-standard/wideband transceivers. Section 2.4 investigates the impact of CMOS technology scaling on linearity, and provides insights into the design of linear LNAs in deep submicron (DSM) processes. Remarks for high linearity LNA design are provided in Section 2.5, and Section 2.6 gives the conclusions.

2.2 Linearization Techniques

A weakly nonlinear amplifier with input X and output Y can be approximated by the first three power series terms:

$$Y = g_1 X + g_2 X^2 + g_3 X^3 \tag{2.1}$$

where $g_{1,2,3}$ are the linear gain and the second/third-order nonlinearity coefficients of the amplifier, respectively. The goal of linearization is to make $g_{2,3}$ small enough to be negligible, keeping only the linear term g_1, hence $Y \approx g_1 X$. The purpose of this chapter is to discuss the main linearization techniques for LNAs.

LNA nonlinearity originates from two major sources:

1) Nonlinear transconductance g_m, which converts linear input voltage to nonlinear output drain current; this effect is also termed “input limited.”

2) Nonlinear output conductance g_{ds}, whose effect becomes apparent under large output voltage swing and small drain-source voltage V_{ds} (i.e. when the device operates near linear region); also referred to as “output limited.”

The MOSFET capacitances (C_{gs}, C_{gd}, C_{db}) are roughly linear when the transistor operates in the saturation region, and when the frequency is less than $f_T/10$ [6]. Their expressions are shown in (2.2):

$$C_{gs} = \frac{2}{3}WLC_{ox} + C_{ol}\,, \quad C_{gd} = C_{ol}\,, \quad C_{db} = C_{jdb} \tag{2.2}$$

where $C_{ol} \approx C_{ox}WL_{overlap}$ the overlap capacitance, and C_{jdb} is is the junction capacitance. Thus, for the most part, the capacitors contribute less distortion than g_m/g_{ds} [44]; however, C_{gd} influences the linearity indirectly through feedback, which will be discussed later.

The IIP3 is degraded by both the intrinsic 3rd-order distortion and the "2nd-order interaction" (caused by intrinsic 2nd-order distortion combined with feedback), while IIP2 originates from intrinsic 2nd-order distortion.

We categorize previously reported CMOS LNA linearization techniques into 8 clusters: a) feedback, b) harmonic termination, c) optimum biasing, d) feedforward, e) derivative superposition (DS), f) IM2 injection, g) noise/distortion cancellation, and h) post-distortion. Note that DS, IM2 injection, and noise/distortion cancellation are special cases of the feedforward technique.

Table 2.1 illustrates the distortion sources and the corresponding linearization methods. Most of the reported linearization techniques focus on suppressing 2nd- and 3rd-order distortion of transconductance. Therefore, linearization of higher order terms (beyond 3rd order) and output conductance still remains an open problem.

2.2.1 Feedback

Fig. 2.1 shows the negative feedback scheme with a nonlinear amplifier A and a linear feedback factor β, where X and Y_c are the input and output signals, respectively. X_f is the feedback signal, and X_e is the difference between X and X_f.

Table 2.1. Overview of distortion sources and linearization techniques

Distortion Sources Linearization Methods	g_m				g_{ds}
	Intrinsic 2^{nd}-order	Intrinsic 3^{rd}-order	2^{nd}-order interaction	Higher order	
Feedback	√	√		√	
Harmonic termination		√	√		
Optimal biasing		√			
Feedforward	√	√		√	
Derivative superposition(DS)		√			
Complementary DS	√	√			
Differential DS	√	√			
Modified DS		√	√		
IM2 injection		√	√		
Noise/distortion cancellation	√	√			√
Post-distortion	√	√			

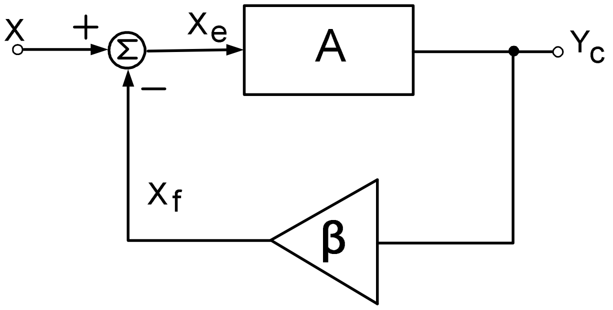


Fig. 2.1. Non-linear amplifier with negative feedback.

Assuming the nonlinear amplifier A can be modeled by (2.1), we obtain the 3rd-order closed-loop power series for Y_c:

$$Y_c = b_1 X + b_2 X^2 + b_3 X^3 \tag{2.3}$$

where the closed loop coefficients related to the open loop coefficients can be derived [see Appendix A]:

$$b_1 = \frac{g_1}{1+T_0}, \quad b_2 = \frac{g_2}{\left(1+T_0\right)^3}, \quad b_3 = \frac{1}{(1+T_0)^4}\left(g_3 - \frac{2{g_2}^2}{g_1}\frac{T_0}{1+T_0} \right) \tag{2.4}$$

where $T_0 = g_1\beta$ is the linear open-loop gain, and $b_{1,2,3}$ are the closed-loop linear gain and second/third-order nonlinearity coefficients, respectively. The IIP2 and IIP3 of the amplifier A and the closed loop system are given by:

$$A_{IIP2,amplifier} = \sqrt{\frac{g_1}{g_2}} \tag{2.5a}$$

$$A_{IIP2,closeloop} = \sqrt{\left|\frac{b_1}{b_2}\right|} = \sqrt{\frac{g_1}{g_2}\left(1+T_o\right)^2} \tag{2.5b}$$

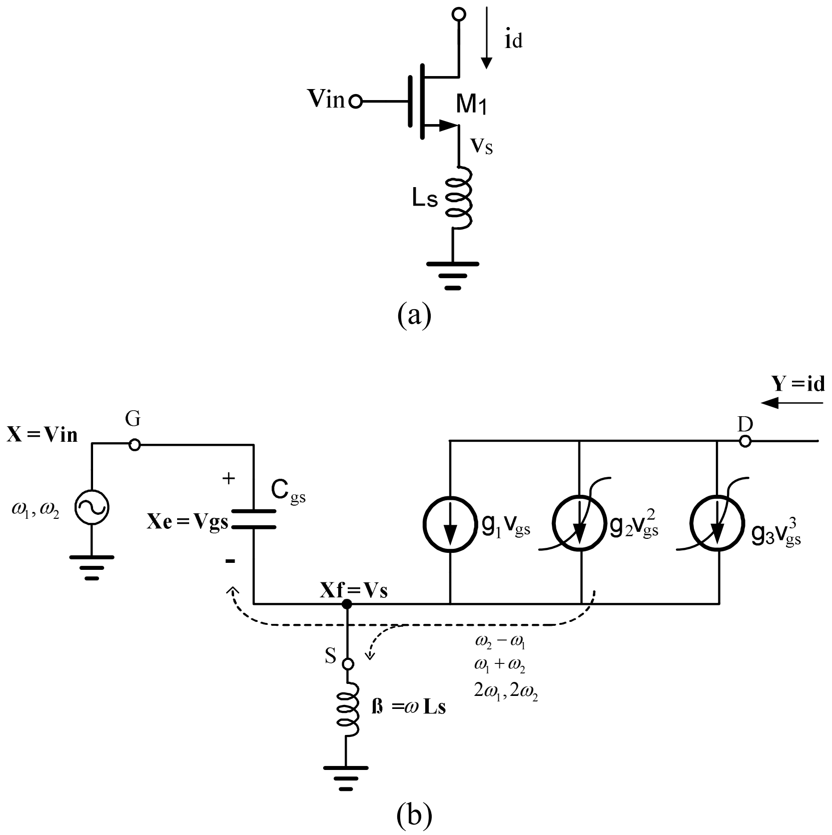
$$A_{IIP3,amplifier} = \sqrt{\frac{4}{3}\left|\frac{g_1}{g_3}\right|} \tag{2.5c}$$

$$A_{IIP3,closeloop} = \sqrt{\frac{4}{3}\left|\frac{b_1}{b_3}\right|} = \sqrt{\frac{4}{3}\frac{g_1}{g_3}\frac{\left(1+T_o\right)^3}{\left(1-\frac{2g_2^2}{g_1 g_3}\frac{T_o}{1+T_o}\right)}} \tag{2.5d}$$

Hence, negative feedback improves A_{IIP2} by a factor of $(1+T_0)$; it also improves A_{IIP3} by a factor of $(1+T_0)^{3/2}$ when $g_2 \approx 0$. As shown by (2.5d), nonzero g_2 degrades IIP3 when g_1 and g_3 have opposite signs (this is the case for typical MOSFET biases). This phenomenon is called "2nd-order interaction" [8]. In other words, whenever the amplifier is connected in feedback, the 3rd-order nonlinearity originates from two sources: 1) intrinsic amplifier 3rd-order nonlinearity. 2) "2nd-order interaction" (originated from intrinsic 2nd-order nonlinearity of the amplifier combined with feedback).

However, feedback linearity improvement is not as effective for LNAs as for baseband circuits because: 1) the open loop gain T_0 cannot be large due to stringent LNA gain, noise, and power requirement. 2) the 2nd-order nonlinearity contributes to the IM3 indirectly through "2nd-order interaction."

To illustrate the "2nd-order interaction," we use the inductively source degenerated LNA [7] as an example. Fig. 2.2 (a) presents the circuit, and Fig. 2.2 (b) shows its small-signal model using the notation from Fig.2.1. $\omega_2 \pm \omega_1$ are the 2nd-order intermodulation components (IM2), and $2\omega_{1,2}$ are the 2nd-order harmonic distortion components. The inductor L_s acts as a frequency-dependent feedback element with $\beta=\omega L_s$, creating a feedback path between the output current i_d and the gate-source voltage v_{in}. For simplicity, we analyze these effects with a Taylor series--for a more accurate, frequency-dependent analysis refer to the results obtained using Volterra series in references [15], [24].



(a)

(b)

Fig. 2.2. (a) Inductively source-degenerated LNA, (b) small-signal model.

First, i_d can be expressed as:

$$i_d = g_1\left(v_{in} - v_s\right) + g_2\left(v_{in} - v_s\right)^2 + g_3\left(v_{in} - v_s\right)^3 \tag{2.6}$$

where g_i is the i^{th}-order coefficient of the nonlinear current of M_1 obtained by taking the derivative of the drain-source DC current I_{DS} with respect to the gate-to-source voltage V_{GS} at the DC bias point:

$$g_1 = \frac{\partial I_{DS}}{\partial V_{GS}},\ \ g_2 = \frac{1}{2!}\frac{\partial^2 I_{DS}}{\partial V_{GS}^2},\ g_3 = \frac{1}{3!}\frac{\partial^3 I_{DS}}{\partial V_{GS}^3} \tag{2.7}$$

When input v_{in} has two frequency components ω_1 and ω_2, we can substitute $v_{in} = A_1 \cos\omega_1 t + A_2 \cos\omega_2 t$ into (2.6) and first assuming $v_s = 0$, we have:

$$
\begin{aligned}
i_d &= g_1\left(A_1\cos\omega_1 t + A_2\cos\omega_2 t\right) \\
&+ g_2\left\{\frac{A_1^2}{2}\left(1+\cos 2\omega_1 t\right) + \frac{A_2^2}{2}\left(1+\cos 2\omega_2 t\right) + A_1 A_2\left[\cos\left(\omega_1+\omega_2\right)t + \cos\left(\omega_1-\omega_2\right)t\right]\right\}^2 \\
&+ g_3\left\{\begin{array}{l}
\dfrac{A_1^3}{4}\left(3\cos\omega_1 t + \cos 3\omega_1 t\right) + \dfrac{A_2^3}{4}\left(3\cos\omega_2 t + \cos 3\omega_2 t\right) \\
+\dfrac{3}{4}A_1^2 A_2\left[\cos\left(2\omega_1 t+\omega_2 t\right) + \cos\left(2\omega_1 t-\omega_2 t\right)\right] + \dfrac{3}{4}A_1 A_2^2\left[\cos\left(\omega_1 t+2\omega_2 t\right) + \cos\left(\omega_1 t-2\omega_2 t\right)\right] \\
+\dfrac{3}{2}A_1^2 A_2\cos\omega_2 t + \dfrac{3}{2}A_1 A_2^2\cos\omega_1 t
\end{array}\right\}
\end{aligned}
\tag{2.8}
$$

Table 2.2 lists the magnitudes for each frequency component:

Table 2.2. Magnitudes and frequency components

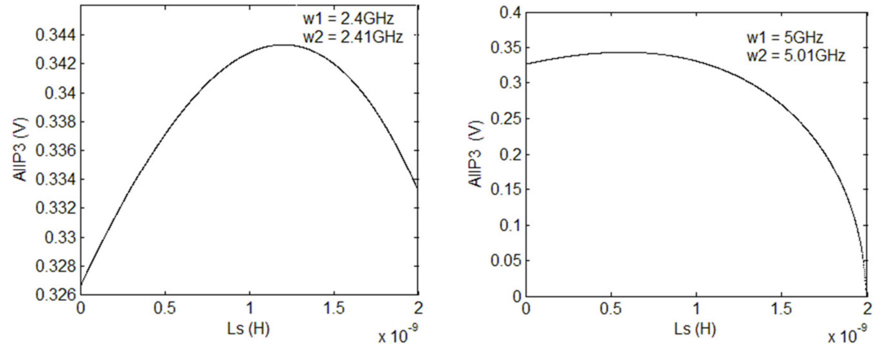
Frequency	Magnitude
ω_1	$g_1 A_1 + g_3\left(\dfrac{3A_1^3}{4} + \dfrac{3}{2}A_1 A_2^2\right)$
ω_2	$g_1 A_2 + g_3\left(\dfrac{3A_2^3}{4} + \dfrac{3}{2}A_1^2 A_2\right)$
$2\omega_1$	$g_2\dfrac{A_1^2}{2}$
$2\omega_2$	$g_2\dfrac{A_2^2}{2}$
$\omega_1 \pm \omega_2$	$g_2 A_1 A_2$
$2\omega_1 \pm \omega_2$	$\dfrac{3}{4}g_3 A_1^2 A_2$
$2\omega_2 \pm \omega_1$	$\dfrac{3}{4}g_3 A_1 A_2^2$

Since $v_s \neq 0$, and it contains components $2\omega_1$, $2\omega_2$, and $\omega_1 \pm \omega_2$ due to the 2nd-order distortion, the product term $-2g_2v_{in}v_s$ from $g_2(v_{in}-v_s)^2$ generates IM3 terms $2\omega_1 \pm \omega_2$ and $2\omega_2 \pm \omega_1$. Therefore, the intrinsic 2nd-order nonlinearity contributes to third-order intermodulation, IM3, when a feedback mechanism is employed. Note that this "2nd-order interaction" problem exists even if the LNA topology is differential because the term $-2g_2v_{in}v_s$ is an odd term and cannot be rejected by differential operation.

Though source degeneration mostly improves linearity, inductive source degeneration actually has two opposing effects on linearity: 1) increases A_{IIP3} by $\approx (1 + g_1\omega Ls)^{3/2}$. 2) Degrades A_{IIP3} due to "2nd-order interaction."

Fig. 2.3 shows A_{IIP3} versus source-degeneration inductor Ls for two cases: input tones at 2.4GHz, 2.41GHz (Fig. 2.3(a)), and at 5GHz, 5.01GHz (Fig. 2.3(b)). Note: 1) this simulation only includes the distortion from input transconductance, while the loading and input- matching resonant network will also contribute to distortion in practice. 2) the gate inductor needs to be adjusted for resonating at different frequencies.

Reducing the degeneration inductance or adding a termination network such that $v_s(\omega) = 0$ at the IM2 frequency can mitigate "2nd-order interaction;" the latter approach is called "harmonic termination."



(a) (b)

Fig. 2.3. A_{IIP3} vs. source-degeneration inductance (a) input tones at 2.4GHz and 2.41GHz (b) input tones at 5GHz and 5.01GHz.

2.2.2 Harmonic Termination

"Harmonic termination" adds a termination network to accomplish one of two effects: 1) BJT case: sets $b_3 = 0$ in (2.4) with the "2nd-order-interaction" term. 2) CMOS case: forces a certain node voltage to zero at the IM2 frequency. This difference is because for BJT, g_1 and g_3 have the same polarity; for CMOS transistors in saturation region, g_1 and g_3 have different polarities; g_1 and g_3 have the same polarity for CMOS transistors in weak inversion region, but the gain becomes very small thus is not practical for LNA design.

Equation (2.4) in Section 2.2.1 was obtained assuming a frequency-independent feedback factor β, which is only valid for pure resistive networks. For frequency-dependent networks as the case in Fig. 2.2, Volterra series [4] should be used to capture the memory effects. To obtain the 3rd-order coefficient in the Volterra series, defined as

$b_3(\omega_x, \omega_y, \omega_z)$, a three-dimensional Fourier transform is performed on the 3rd-order impulse response $h_3(\tau_x, \tau_y, \tau_z)$ of the system. Thus, (2.4) becomes [see Appendix B.5]:

$$b_3\left(\omega_x,\omega_y,\omega_z\right)=\frac{1}{\left(1+T\left(\omega_x+\omega_y+\omega_z\right)\right)\left(1+T\left(\omega_x\right)\right)\left(1+T\left(\omega_y\right)\right)\left(1+T\left(\omega_z\right)\right)}\times$$
$$\left[g_3-\frac{2g_2^{\ 2}}{3g_1}\left(\frac{T\left(\omega_y+\omega_z\right)}{1+T\left(\omega_y+\omega_z\right)}+\frac{T\left(\omega_x+\omega_z\right)}{1+T\left(\omega_x+\omega_z\right)}+\frac{T\left(\omega_x+\omega_y\right)}{1+T\left(\omega_x+\omega_y\right)}\right)\right] \tag{2.9}$$

$T(\omega)=g_1\beta(\omega)$ is the frequency-dependent linear loop gain, which also depends on the feedback components and termination impedances Z_1, Z_2, and Z_3 shown in Fig. 2.4.

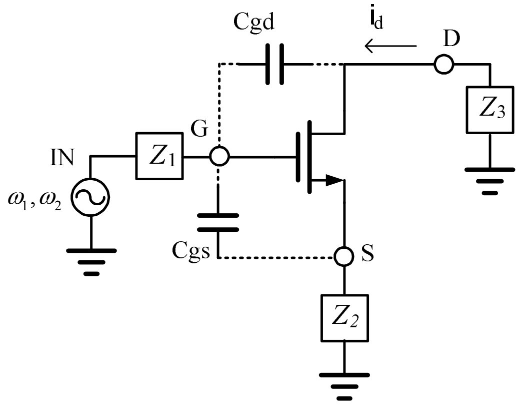


Fig. 2.4. Common-source LNA with termination impedances.

The expressions $|b_3(\omega_x, \omega_y, \omega_z)|$ and $\angle b_3(\omega_x, \omega_y, \omega_z)$ give the magnitude and phase of a tone at frequency $\omega_x+\omega_y+\omega_z$ generated by 3rd-order nonlinearity. For example, given two input tones at ω_1 and ω_2, to get the IM3 products at $2\omega_1-\omega_2$, choose $\omega_x = \omega_y = \omega_1$ and $\omega_z = -\omega_2$. Assuming two closely spaced tones, i.e. $-\omega_1=\omega$, $\omega_2 = -\omega-\Delta\omega$, and$\Delta\omega \approx 0$, we have

$$b_3\left(\omega,\omega,-\omega-\Delta\omega\right) \cong \frac{1}{\left(1+T\left(\omega\right)\right)^3\left(1+T\left(-\omega\right)\right)} \times \left[\underbrace{g_3 - \underbrace{\frac{2g_2^{\;2}}{3g_1}\left(\frac{2T\left(\Delta\omega\right)}{1+T\left(\Delta\omega\right)} + \frac{T\left(2\omega\right)}{1+T\left(2\omega\right)}\right)}_{A_2}}_{\varepsilon\left(\Delta\omega,2\omega\right)} \right] \tag{2.10}$$

From (2.10), the contribution of 2^{nd}-order distortion to IM3 is defined by the loop gain at sub-harmonic frequency Δω and 2^{nd}-harmonic frequency 2ω, i.e. T(Δω) and T(2ω). Therefore, by tuning the termination impedances at Δω and/or 2ω, the amplitude and phase of the 2^{nd}-order interaction terms A_2 can be adjusted to cancel the intrinsic 3^{rd}-order distortion term g_3, so that $b_3 \approx 0$. For narrowband applications, Δω and 2ω are usually out-of-band, keeping the in-band operation unaffected, hence the "harmonic termination" technique is also called "out-of-band tuning/termination" [9], [10].

The 2^{nd}-order nonlinear current can mix with the input through three intrinsic feedback paths, as listed in Table 2.3 for the common source LNA(CS-LNA) and common gate LNA(CG-LNA):

Table 2.3. Three intrinsic feedback paths

Feedback Path	Path Components	
	CS-LNA	CG-LNA
Source-to-gate	C_{gs} + source degeneration inductor Z_2	C_{gs} + input driving impedance [11]
Drain-to-gate	C_{gd} + output load Z_3	
Input-to-gate	Input matching network Z_1	

The CG-LNA inherently has less drain-to-gate feedback than the CS-LNA since its gate is AC grounded, therefore the CG-LNA usually has better linearity.

Resonant tanks can be added to optimally tune $Z_i(\Delta\omega)$ and/or Z_i (2ω) (i=1-3) such that the 2nd-order remixing term cancels the IM3 term. Techniques have been reported to tune the input terminal $Z_1(\Delta\omega)$ for bipolar LNAs [9]-[11]. The terminations are commonly implemented with dedicated LC networks, which provide high impedance at ω but small impedance paths to ground at $\Delta\omega$ or 2ω. However, the required inductance value is usually quite large. The low Q factors of on-chip passive inductors limit their distortion-cancellation effectiveness and also affect noise and input matching. Furthermore, on-chip active inductors add noise and nonlinearity, so in practice off-chip inductors are employed.

Though popular in BJT LNAs, harmonic termination is less effective for CMOS LNAs [10], [13]. For a stable design, the A_2 term in (2.10) has a positive real part. Thus, $|\varepsilon(\Delta\omega,2\omega)|$ can be reduced below $|g_3|$ only if g_3 is positive, which is true for a BJT, but not for a MOSFET in saturation. Therefore, both g_3 and A_2 must be reduced to improve a CMOS LNA's IIP3.

From (2.10), one way to reduce A_2 is to reduce both $Z_i(\Delta\omega)$ and Z_i (2ω) (i=1-3) [14]. A cascode configuration can reduce Z_3 to $1/g_1$ [14], and capacitive cross-coupling in the cascode stage further reduces Z_3 to $1/(2g_1)$ [15]. Although their IIP3 improvement is not as great as that attainable from large passive LC components, it is more feasible. In [12] and [14], an LC-resonant RF current source reduces Z_2. Fig. 2.5 shows some termination examples, in which L_t and C_t form low-frequency/2nd-harmonic trap networks Z_1 ($\Delta\omega$) (Fig. 2.5 (a)) and $Z_2(\Delta\omega, 2\omega)$ (Fig. 2.5(b)).

In Fig. 2.5(b), IIP3 can be expressed in Volterra series as [12]:

$$IIP3 = \frac{g_1}{6} \frac{1}{\left| \left[1 - g_1/\left(g_1 + Y(\omega)\right) \right] \cdot \left| g_1 + Y(\omega) \right|^2 K(\omega) \right|} \times R_s \tag{2.11}$$

$$K(\omega) = g_3 + \frac{2}{3} g_2^2 \left[\frac{1}{Y(2\omega) + g_1} + \frac{2}{Y(\Delta\omega) + g_1} \right] \tag{2.12}$$

where Y(ω) is the admittance at the transistor source, and R_s is the signal source resistance. The parallel LC network helps to increase Y(2ω) and Y(Δω), minimizing the “2nd-order interaction” and improving IIP3.

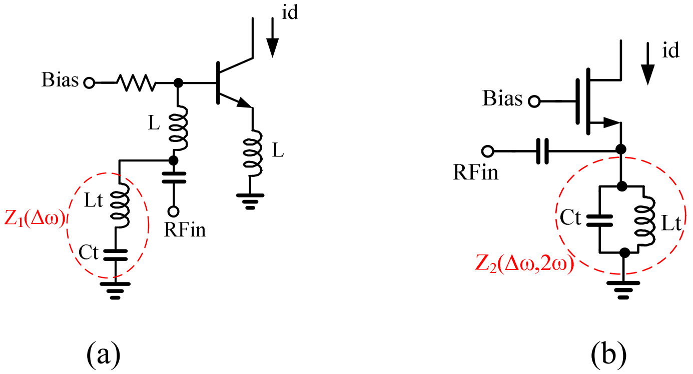


Fig. 2.5. Harmonic termination: (a) common-emitter stage with low-frequency-trap network (L is the package inductance) [11], (b) common-gate stage with RF current source [14].

Harmonic termination only works well in narrowband systems because the tuning network is optimized at Δω and 2ω, and only works for a narrow range of two-tone spacing/center frequencies [9]. For wideband applications, Δω and 2ω vary considerably, so it is difficult to tune out the termination impedance. Furthermore, Δω and 2ω may fall in-band, affecting the normal operations.

In summary, to improve CMOS LNA linearity, we should ensure a small intrinsic 3rd-order coefficient g_3 of the transistor, and relax the "2nd-order interaction." Adding a harmonic termination network alleviates the latter. Next, we will discuss a few techniques to reduce the third-order coefficient g_3.

2.2.3. Optimal Biasing

Assume the main nonlinearity of a MOS transistor arises from transconductance nonlinearity, as modeled in (2.6). To characterize this single-transistor nonlinearity, we fixed its drain-source voltage V_{ds}, swept the gate-source voltage V_{gs}, and then took the first three derivatives of the drain-source DC current I_{ds} with respect to V_{gs} (as defined in Equation (2.7)) at every DC bias point to obtain the plots in Fig. 2.6. If we define the inversion level of the transistors as: $i_f = I_{ds}/I_s$, where $I_s = 0.5\ \mu Cox\Phi_t^2 W/L$ is the normalized current, then i_f= 0-960 in these plots for the chosen V_{gs} sweep range.

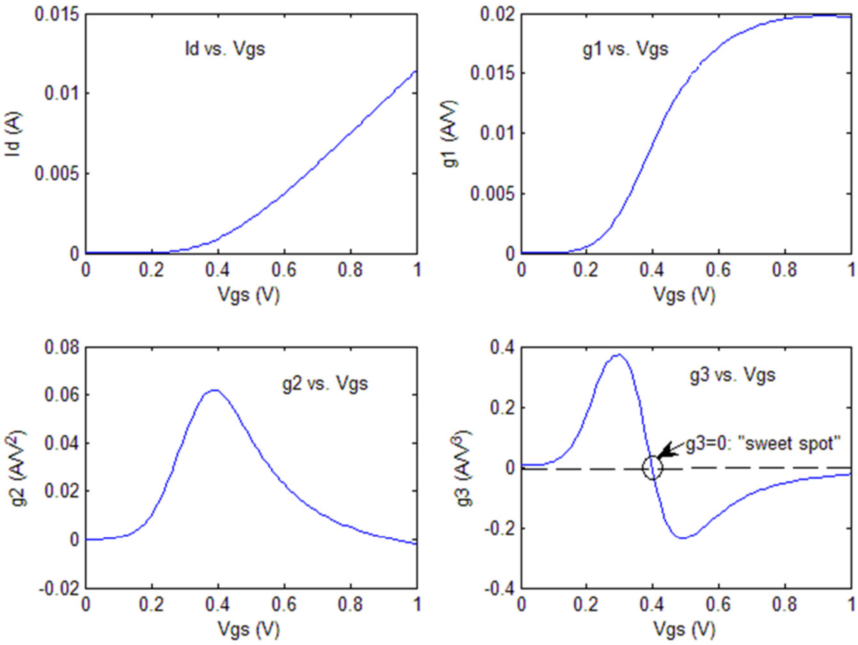


Fig. 2.6. NMOS transconductance characteristics
(UMC 90nm CMOS process, W/L = 20/0.08μm, V_{ds} = 1V).

While g_2 is always positive, g_3 has a sign inversion:

- Small V_{gs}: $g_3 > 0$ because the transistor operates in weak inversion, where the I_{ds} vs V_{gs} relation is exponential.

- Large V_{gs}: $g_3 < 0$ because mobility degradation/velocity saturation cause gain compression. The key idea of "optimum biasing" is to bias the transistor at the "sweet spot" g_3=0 [6], which is the "moderate inversion" region. The inversion level is 24 at the "sweet spot" with our specific biasing and sizing.

Though simple in principle, the optimal biasing technique has the following limitations:

1) The cancellation is sensitive to process variations (e.g. V_{th}), therefore, it is recommended to use constant-current or constant-gm biasing over constant-voltage biasing.

2) The technique is sensitive to operating point, resulting in a limited input-signal amplitude range for effective distortion cancellation.

3) The sweet spot shifts to a lower bias current level as the gain increases, since the output swing increases and nonlinear output conductance starts to play a role.

4) Due to the "2nd-order interaction," the IIP3 peak at the "sweet spot" decreases and will finally disappear as source degeneration inductance increases.

5) The sweet spot is frequency-dependent, and the IIP3 peak decreases due to parasitic effects [6].

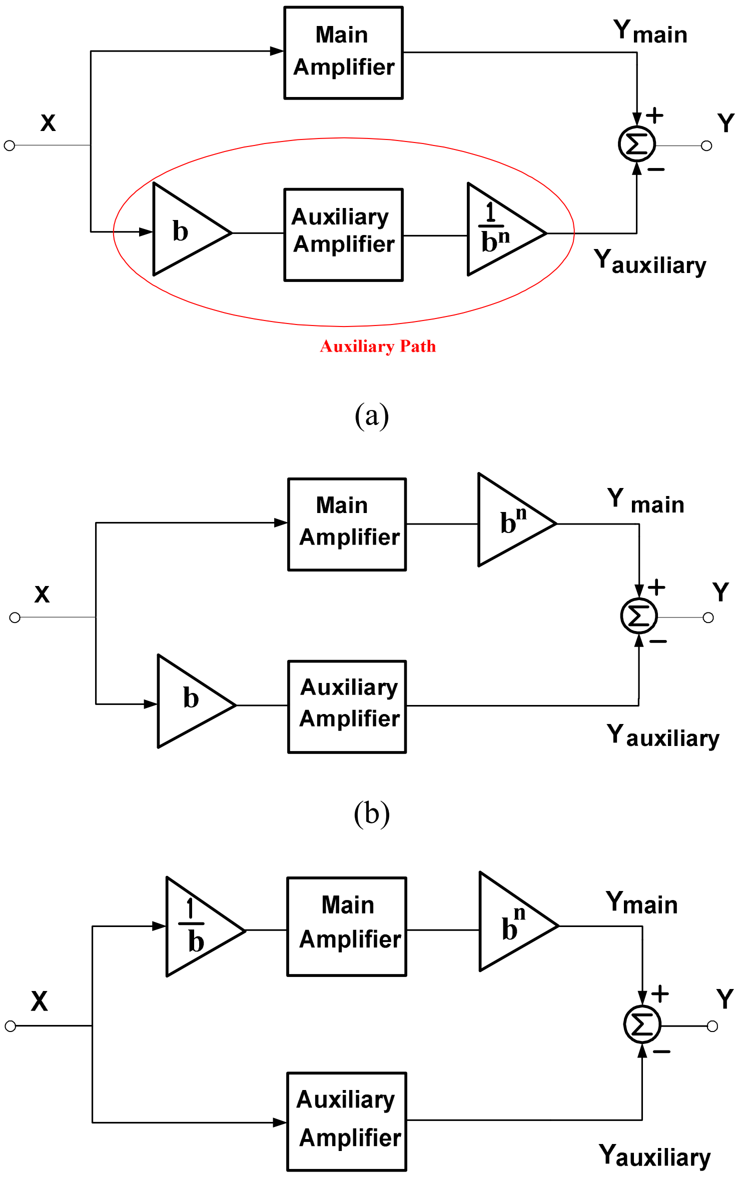
6) Biasing the transistor at $g_3 = 0$ restricts the input-stage transconductance, lowering gain and increasing NF.

An automatic bias circuit could mitigate some of these effects [9]; however, this "automatic" bias circuit itself is prone to process variations and requires manual tuning in practice. The bias point for optimum IIP3 is shifted from the bias for zero g_3 due to "2nd-order interaction."

In summary, the "sweet spot" is a single transistor characteristic and only signifies optimum intrinsic 3rd-order transconductance nonlinearity. Many other factors will weaken the IIP3 improvement at the "sweet spot". Furthermore, some claim that no "sweet spot" exists in practical LNAs because of input/output networks and parasitics[6].

2.2.4 Feedforward

From equation (2.5), note that simultaneous cancellation of g_2 and g_3 with minimum effects on g_1 requires more degrees of freedom. Generating additional nonlinear currents/voltages, and subsequently summing (subtracting) them accomplishes such simultaneous cancellation. These actions constitute feedforward, as illustrated in Fig. 2.7(a) [16]. An auxiliary path includes a replica amplifier and signal-scaling factors b and $1/b^n$ at its input/output, respectively, to replicate the distortion in the main path. We use n = 2 or 3 depending on whether IM2 or IM3 is to be cancelled. Note that if the amplifiers are differential, the 2nd-order distortion is ideally zero and n = 3 yields a linear output. Without loss of generality, the following discusses the single-ended case. To obtain the total output Y, we subtract the output of the auxiliary amplifier ($Y_{auxiliary}$) from that of the main amplifier (Y_{main}). Assuming $|b|>1$, by changing the location and value of scaling factor, we propose two alternate implementations, shown in Fig. 2.7(b) and (c), respectively.



(a)

(b)

(c)

Fig. 2.7. Three representations of the feedforward linearization technique.

Assuming the main and auxiliary amplifiers have the same nonlinearity coefficients g_i, we have,

(a)

$$Y_{main} = g_1X + g_2X^2 + g_3X^3 \tag{2.13}$$

$$Y_{auxiliary} = \left[g_1\left(bX\right) + g_2\left(bX\right)^2 + g_3\left(bX\right)^3\right]\frac{1}{b^n} \tag{2.14}$$

$$Y = Y_{main} - Y_{auxiliary} = g_1\left(1-\frac{1}{b^{n-1}}\right)X + \underbrace{g_2\left(1-\frac{1}{b^{n-2}}\right)X^2 + g_3\left(1-\frac{1}{b^{n-3}}\right)X^3}_{\text{Residue Distortion}} \tag{2.15}$$

(b)

$$Y_{main} = \left(g_1X + g_2X^2 + g_3X^3\right)b^n \tag{2.16}$$

$$Y_{auxiliary} = g_1\left(bX\right) + g_2\left(bX\right)^2 + g_3\left(bX\right)^3 \tag{2.17}$$

$$Y = g_1b\left(b^{n-1}-1\right)X + \underbrace{g_2b^2\left(b^{n-2}-1\right)X^2 + g_3b^3\left(b^{n-3}-1\right)X^3}_{\text{Residue Distortion}} \tag{2.18}$$

(c)

$$Y_{main} = \left[g_1\frac{X}{b} + g_2\left(\frac{X}{b}\right)^2 + g_3\left(\frac{X}{b}\right)^3\right]b^n \tag{2.19}$$

$$Y_{auxiliary} = g_1X + g_2X^2 + g_3X^3 \tag{2.20}$$

$$Y = g_1\left(b^{n-1}-1\right)X + \underbrace{g_2\left(b^{n-2}-1\right)X^2 + g_3\left(b^{n-3}-1\right)X^3}_{\text{Residue Distortion}} \tag{2.21}$$

where $g_2 = 0$ for differential amplifiers. Comparing (2.15), (2.18), and (2.21), the implementation in (a) has a gain-attenuation factor of $(1\text{-}1/b^{n\text{-}1})$, thus gain is reduced by 2.5dB with b = 2 and n = 3 as in [17]. On the other hand, the proposed implementations in (b) and (c) increase the gain. Note that (c)'s input attenuator 1/b worsens its NF. The implementations in Fig. 2.7 can only cancel one type of harmonic at a time; to reduce both 2nd- and 3rd-order distortion simultaneously, we need an additional degree of freedom, which we could attain with two auxiliary paths as shown in Fig. 2.8.

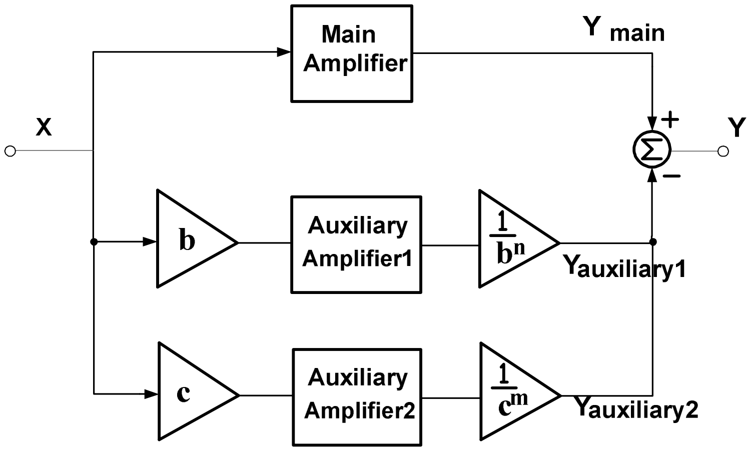


Fig. 2.8. Proposed dual-auxiliary-path feedforward linearization technique.

Assuming the main and auxiliary amplifiers have the same nonlinearity coefficients g_i, we have:

$$Y_{auxiliary1} = \left[g_1(bX) + g_2(bX)^2 + g_3(bX)^3\right]\frac{1}{b^n} \tag{2.22}$$

$$Y_{auxiliary2} = \left[g_1(cX) + g_2(cX)^2 + g_3(cX)^3\right]\frac{1}{c^m} \tag{2.23}$$

$$\begin{aligned} Y &= Y_{main} - Y_{auxiliary1} - Y_{auxiliary2} \\ &= \underbrace{g_1\left(1 - \frac{1}{b^{n-1}} - \frac{1}{c^{m-1}}\right)X}_{\text{Linear Gain}} + \underbrace{g_2\left(1 - \frac{1}{b^{n-2}} - \frac{1}{c^{m-2}}\right)X^2 + g_3\left(1 - \frac{1}{b^{n-3}} - \frac{1}{c^{m-3}}\right)X^3}_{\text{Residue Distortion}} \end{aligned} \tag{2.24}$$

In (2.24), we have two equations (2nd and 3rd term equals zero) and four variables (b, c, n, m), resulting in multiple solutions. A possible additional constraint is to bound the reduction in linear gain to be less than, say, 20%, and one reasonable solution set is: b = -2, c = -3, n = 0, m = 1. This choice causes the linear gain to double. For a more general case, if we specify that the g1 is scaled by K after linearization as in (2.21), we can choose reasonable values for m and n, then obtain the values for b and c as follows:

$$1-\frac{1}{b^{n-1}}-\frac{1}{c^{m-1}}=K \tag{2.25}$$

$$Kc^{m-1}+\left(1-K\right)c^{2}-2c+1=0 \tag{2.26}$$

$$b=\frac{c^{\frac{m-1}{n-1}}}{\left[\left(1-K\right)c^{m-1}-1\right]^{\frac{1}{n-1}}} \tag{2.27}$$

Note that if the amplifiers are differential, all even order harmonics are ideally zero, and the implementation in Fig. 2.8 can cancel both 3rd- and 5th-order distortion.

This general feedforward technique improves linearity without relying on the amplifier's linearity characteristics; however, it has several disadvantages:

1) Accurate, noiseless, and highly linear scaling factors (b, c) are often not feasible. For instance, the off-chip coaxial assembly used in [17] is expensive and cannot be integrated.

2) The added active components introduce more noise.

3) Highly sensitive to mismatch between the main and auxiliary gain stages.

4) Large power overhead due to the auxiliary amplifier. In worst case, the auxiliary amplifier is an exact copy of the main amplifier, thus the power is doubled or tripled. [18] reports an improved feedforward technique, where the auxiliary path only

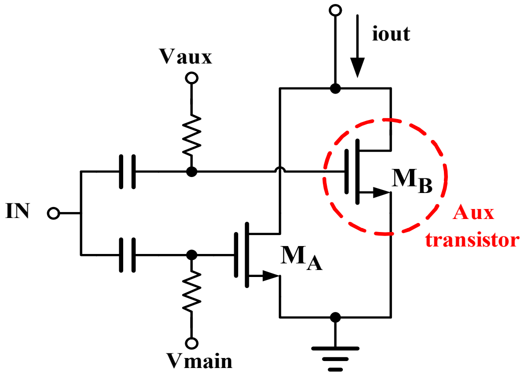
passes the IM3 products. Hence, its dynamic range is relaxed, resulting in only 21% power overhead.

Next, we will discuss three special cases of the feedforward technique: derivative superposition, IM2 injection, and noise/distortion cancellation.

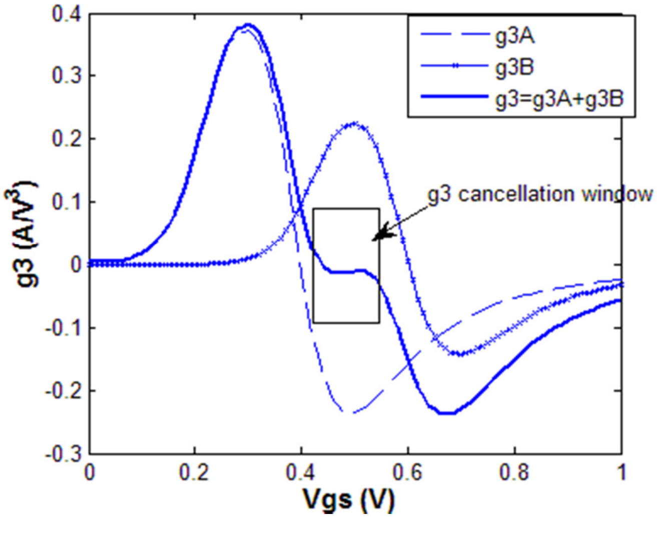
2.2.5 Derivative Superposition (DS)

The “Derivative Superposition (DS)” method [12], [14], [19]-[21] is a special case of the feedforward technique. Notice that the DS method is obtained when b=1 in Fig. 2.7 and when the main/auxiliary amplifiers are implemented with transistors operating in different regions. Fig. 2.9(a) depicts a dual-NMOS implementation of the DS method. $M_{A/B}$ denotes the main/auxiliary transistor, respectively, and the input matching network is omitted for simplicity.

This method is called “derivative superposition” because it adds the 3rd derivatives (g_3) of drain current from the main and auxiliary transistors to cancel distortion. As discussed in Section 2.2.3, g_3’s sign changes at the boundary of moderate and strong inversion region. Thus, proper biasing creates net zero g_3, as shown in Fig. 2.9(b). Linearity is improved within a finite bias-voltage *range* instead of just a *point*. Hence the DS method is less sensitive to process variations than the optimum biasing technique. Moreover, the auxiliary path contains only one weak-inversion transistor, resulting in much smaller power consumption than the general feedforward technique. Since the DS method employs multiple transistors in parallel with their gates connected together, it is also called the “multiple gated transistor technique" (MGTR) [12], [14]. Note that since positive and negative characteristic of g_3 are not symmetric, the g_3- cancellation window is fairly narrow with only one auxiliary transistor, but the window widens with more auxiliary transistors at the cost of degraded input matching, NF, and gain [20].



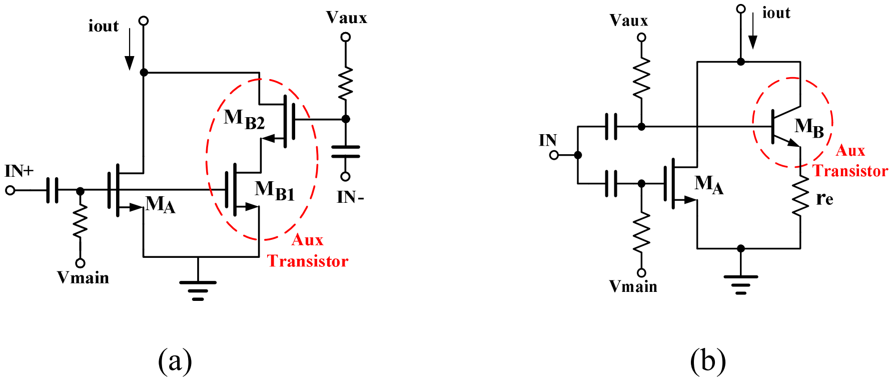
(a)



(b)

Fig. 2.9. (a) DS method with dual-NMOSs, (b) 3rd order distortion terms of the main transistor (g_{3A}), auxiliary transistor (g_{3B}), and total output (g_3) (UMC 90nm CMOS process, $(W/L)_{MA}$ = 20/0.08μm, $(W/L)_{MB}$ = 12/0.08μm, V_{ds} = 1V).

Fig. 2.10(a) and (b) show alternate implementations of the DS method that use a CMOS transistor in triode region [19] or bipolar [20] transistor as the auxiliary device. In Fig. 2.10(a), M_{B1} and M_{B2} are driven by differential input signals. M_{B1} is biased in deep triode region, and M_{B2} helps to boost the positive g_3 peak of M_{B1} to be sufficiently large to cancel the negative peak in g_3 of input transistor M_A. In Fig. 2.10(b), a bipolar transistor M_B provides the positive g_3, and emitter degeneration resistor r_e reduces g_3 to match that of M_A for optimum distortion cancellation.



(a) (b)

Fig. 2.10. DS method: (a) additional transistor works in triode region [19], (b) use of a bipolar transistor [20].

2.2.5.1 Complementary DS

Fig. 2.6 shows that the 2nd-order term (g_2) has a positive sign for transistors working in either moderate or strong inversion region. Therefore, techniques, such as conventional DS, that improve 3rd-order distortion usually worsen 2nd-order distortion.

The “Complementary DS method” employs an NMOS/PMOS pair to improve IIP3 without hurting IIP2 [22], [29].

Fig. 2.11 shows the common-source and common-gate implementations, respectively. The AC current combiner in Fig. 2.11(b) could be seen either as a large coupling capacitor (e.g. 15pF in [29]) with negligible impedance within signal bandwidth, or as a current mirror [22]. Since the AC input signal for NMOS/PMOS are out of phase, the output current is expressed as:

$$i_{dsn} = g_{1A} v_{gs} + g_{2A} v_{gs}^2 + g_{3A} v_{gs}^3 \tag{2.28}$$

$$i_{dsp} = -g_{1B} v_{gs} + g_{2B} v_{gs}^2 - g_{3B} v_{gs}^3 \tag{2.29}$$

$$i_{out} = i_{dsn} - i_{dsp} = \left(g_{1A} + g_{1B}\right) v_{gs} + \left(g_{2A} - g_{2B}\right) v_{gs}^2 + \left(g_{3A} + g_{3B}\right) v_{gs}^3 \tag{2.30}$$

The total transconductance increases, the IM2 term decreases because g_{2A} and g_{2B} have the same sign, and the IM3 term decreases because g_{3A} and g_{3B} have different signs. Fig. 2.12 compares the conventional DS and complementary DS in terms of 2nd-order (g_2) and 3rd-order (g_3) distortion of the output current. A cancellation window for g_3 exists in both cases at V_{gs} around 500mV, but g_2 is maximized for conventional DS and minimized for complementary DS. Note that the g_3 cancellation window is narrower and less flat for complementary DS since PMOS and NMOS devices have different linearity characteristics, so the IIP3 improvement is not as good as that in a dual-NMOS implementation. Furthermore, as shown in (2.30), we can either match g_{3A} and g_{3B} for a good IIP3 while slightly cancelling g_2, or we can match g_{2A} and g_{2B} for optimum IIP2, because IIP2 and IIP3 do not share the same optimum bias. The differential DS method

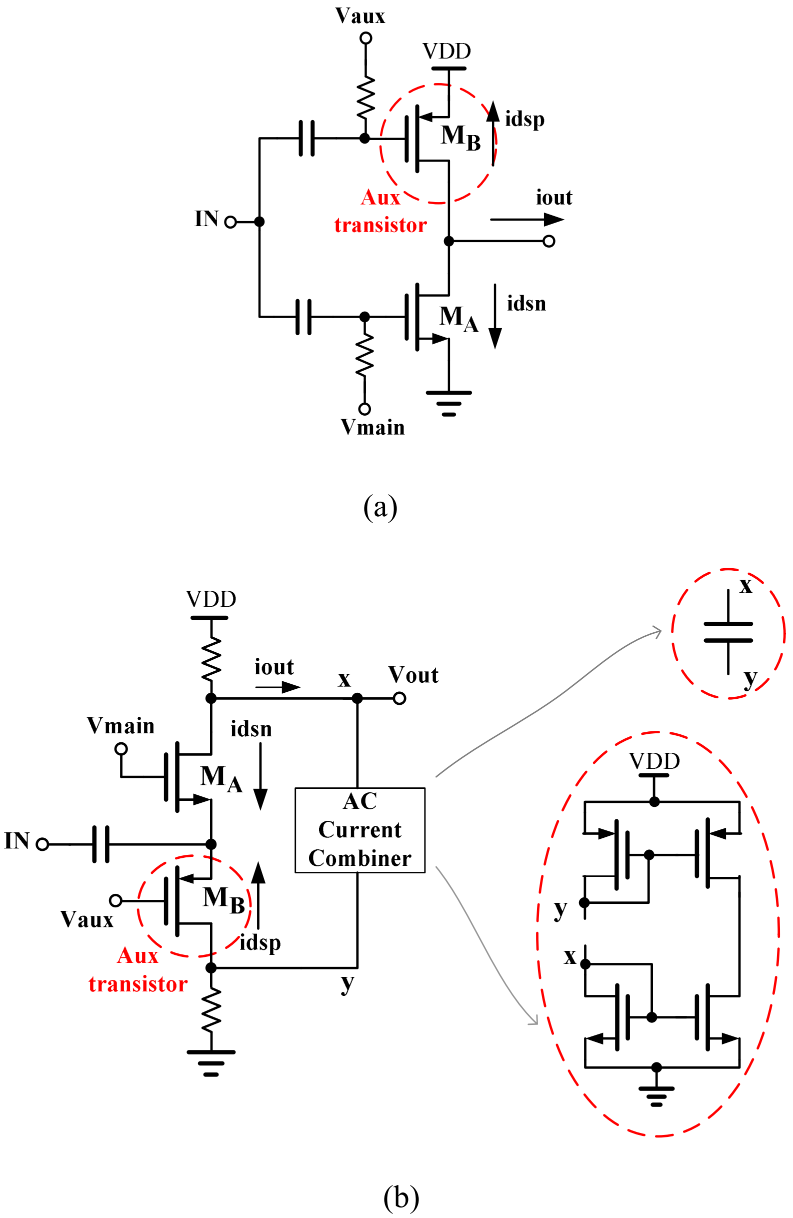
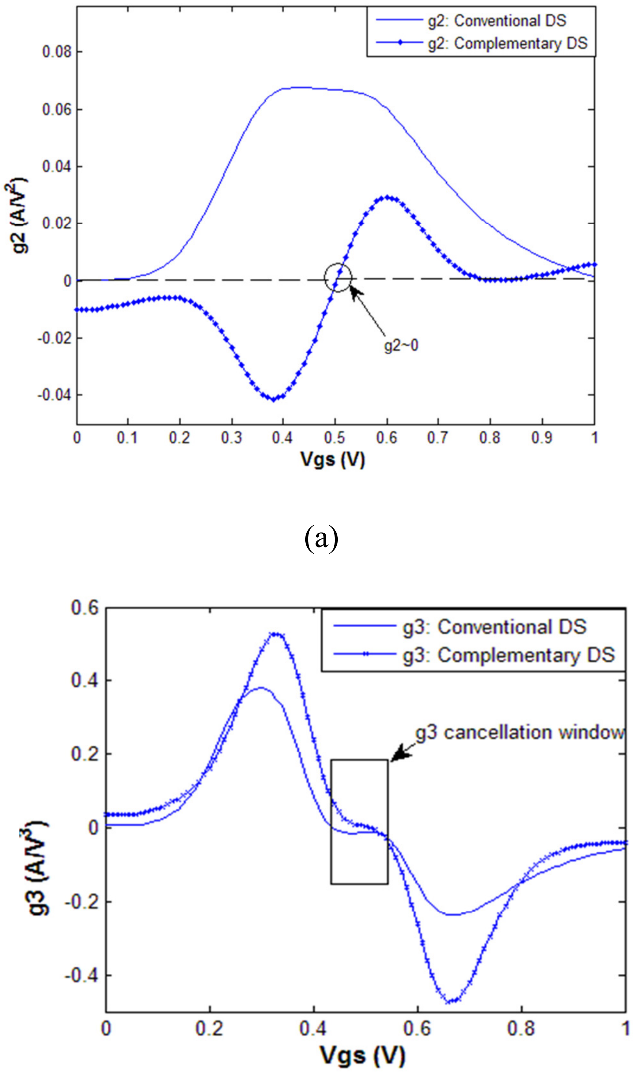


Fig. 2.11. (a) Complementary DS with common-source configuration [22], (b) complementary DS with common-gate configuration [29].

is essentially the same as complementary DS, which also alleviates IIP2 problem [20], [23].

As illustrated in equations (2.4), (2.9) and (2.10), the "2nd-order interaction" ultimately limits the IIP3 improvement at higher frequencies after the intrinsic g_3-induced 3rd-order distortion is cancelled by the DS method. The “Modified DS method” alleviates this issue [24], [25].



(a)

(b)

Fig. 2.12. Comparison of conventional (dual-NMOS) DS and complementary (PMOS/NMOS) DS: (a) g_2 vs. V_{gs} (b) g_3 vs. V_{gs} (UMC 90nm CMOS process, V_{ds} = 1V).

2.2.5.2 Modified DS

As discussed in Section 2.2.2, three feedback paths exist for "2nd-order interaction": source-to-gate, drain-to-gate, and input-to-gate. The modified DS methods [24], [25] provide an on-chip solution to minimize the source-to-gate feedback.

The vector diagram in Fig. 2.13 graphically explains the modified-DS concept, and Fig. 2.14(a) shows the circuit implementation [24]. Note that choice of L_2 determines the angle of g_{3B}. In conventional DS, as illustrated in Fig. 2.13(a), equations (2.31) and (2.32), the anti-parallel g_{3A} and g_{3B} result in a zero total g_3, but residual IM3 exists due to g_{2A} contributions (Note: here we neglect g_{2B}). In the modified DS method, shown in Fig. 2.13(b), equations (2.33) and (2.34), g_{3B} is rotated properly such that the composite vector of g_{3A} and g_{3B} contribution is 180° out of phase with the g_{2A} contribution, yielding zero net IM3.

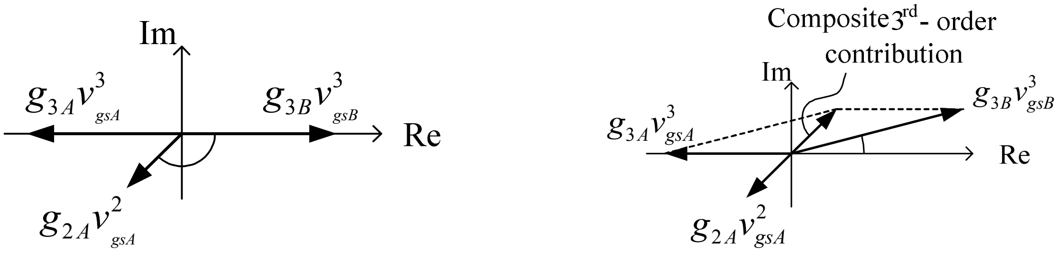
$$IIP3_{conventional} = \frac{4g_1^2\omega^2 LC_{gs}}{3\left|\varepsilon_{conventional}\right|} \tag{2.31}$$

$$\varepsilon_{conventional} = g_3 - \frac{2g_2^2/3}{g_1 + \frac{1}{j2\omega L} + j2\omega C_{gs} + Z_1(2\omega)\frac{C_{gs}}{L}} \tag{2.32}$$

$$IIP3_{\text{mod}ified} = \frac{4g_{1A}^2\omega^2\left[L_1\left(C_{gsA}+C_{gsB}\right)+L_2C_{gsA}\right]}{3\left|\varepsilon_{\text{mod}ified}\right|} \tag{2.33}$$

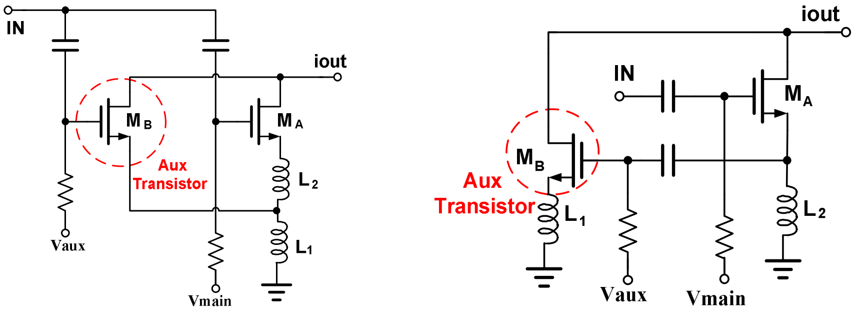
$$\begin{aligned}\varepsilon_{\text{mod}ified} &= g_{3B}\left(1+j\omega L_2 g_{1A}\right)\left[1+\left(\omega L_2 g_{1A}\right)^2\right] \\ &\times\left[1+\frac{L_2C_{gsA}}{L_1\left(C_{gsA}+C_{gsB}\right)+L_2C_{gsA}}\right]+g_{3A}-\frac{2g_{2A}^2}{3g_{1A}}\frac{1}{1+\frac{1}{j2\omega\left(L_1+L_2\right)g_{1A}}}\end{aligned} \tag{2.34}$$

Although the channel noise of weak inversion transistor M$_B$ is negligible, its gate-induced noise is inversely proportional to drain current, and is added directly to the main transistor's (M$_A$) gate noise because their gates are connected together. M$_B$ also affects the input impedance matching. An alternate implementation of the modified DS method reported in [25] (Fig. 2.14(b)) moves M$_B$ to the source of M$_A$ instead of directly connecting it to the input, thus minimizing the degradation in NF and input matching.



(a) (b)

Fig. 2.13. Vector diagram for the distortion components of (a) conventional DS method (b) Modified DS method [24].



(a) (b)

Fig. 2.14. Circuit implementation of modified DS method (a) [24] (b)[25].

Limitations of the DS methods include the following:

1) The weak-inversion transistor may not operate at sufficiently high frequency.

2) The weak-inversion transistor cannot handle large signals or it will be turned off, resulting in a very limited distortion-cancellation range.

3) Weak-inversion transistor models are generally not accurate, resulting in considerable discrepancy between simulation and measurement.

4) Triode-region transistors' positive g_3 peaks decrease as technology scales down, thus complicating the task of matching their amplitudes with the negative peaks of g_3 in main transistors.

5) Matching transistors working in different regions or matching bipolar with MOS transistors is difficult if not impossible, resulting in a linearity improvement sensitive to PVT variations. Current bias with digital control bits [14] or manual adjustment is required for good results in practice.

Fig. 2.15 shows an example IIP3 measurement plot [19], while the corresponding circuit has been shown in Fig. 2.10(a). Although it is from a conventional DS method, similar characteristics can be observed with complementary, differential, and modified DS methods. We observe the following:

1) The DS method works well within the g_3-cancellation window annotated in Fig. 2.9(b) and Fig. 2.12 (P_{in} < -20dBm).

2) Even for inputs outside the g_3-cancellation window, the DS method can still reduce the 3rd-order tone below that of the conventional LNA having a main transistor with negative g_3 as long as g_3 of the auxiliary transistor is positive.

3) The 3rd-order curve shows a greater-than-three slope at much smaller input amplitudes after applying the DS method, because the 5th and higher odd-order-distortion terms contribute more appreciably after g_3 is cancelled.

4) The DS method does not improve the compression point, because it is effective for a small input signal, while the compression point quantifies large signal behavior.

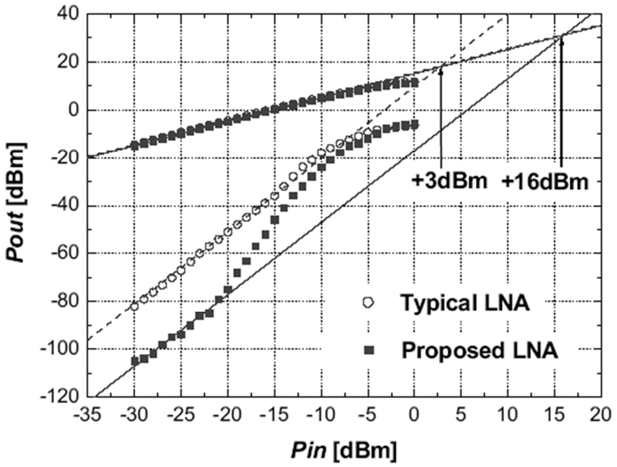


Fig. 2.15. Measured IIP3 of LNAs with/without DS method [19] (© 2003 IEEE).

2.2.6. IM2 Injection

The IM2 Injection method eliminates the explicit auxiliary path entirely by merging it with the main path to reuse the active devices and the DC current [26]. To understand the concept, we first recall equation (2.10): to reduce IM3, we should minimize the $\varepsilon(\Delta\omega,2\omega)$ term. As previously discussed in Section 2.2.2, making A_2 cancel g_3 is difficult for CMOS LNAs because these two terms are out-of-phase in a typical design. Furthermore, the self-generated IM2 term has too small amplitude to suppress g_3 sufficiently.

The IM2 injection technique externally generates and injects a low-frequency IM2 component into the circuit. The injected IM2 phase is inverted with boosted amplitude for IM3 cancellation. Hence, IM2 injection could also be viewed as a smart implementation of harmonic termination. Fig. 2.16 illustrates the concept and basic cells. M1 and M2 are the input transistors of the LNA, and M4, M5, R, and C compose a squaring circuit to generate a low-frequency IM2 current at $\omega_2 - \omega_1$, which is then injected through M3 into the common source node v_s of the LNA. This technique utilizes 2nd-order interaction to generate tones at $2\omega_2 - \omega_1$ and $2\omega_1 - \omega_2$ to cancel the IM3 tones arising from intrinsic 3rd-order distortion. Detailed derivations can be found in [26], the key idea is to match the amplitude of the IM2 current from the squaring circuits and the main circuit for optimal distortion cancellation, and the design equation is:

$$\underbrace{-\frac{2g_{1,M1}g_{3,M1}}{4g_{2,M1}}+\frac{3}{2}g_{2,M1}}_{\text{Main Circuit}}=\underbrace{-g_{1,M3}\times 2g_{2,M4}R}_{\text{Squaring Circuit}} \tag{2.35}$$

where $g_{i,Mi}$ is the i_{th} transconductance coefficient of M_i. The injected IM2 tone should be in phase with the envelope of the RF input signal. Because it is easier to match the phase at low frequency, frequency component $\omega_2 - \omega_1$ is chosen over other IM2 components ($\omega_2 + \omega_1$, $2\omega_2$, $2\omega_1$), by adjusting the bandwidth of the RC filter. Since the linear gain is added in phase, and the noise injected from the IM2 generator appears as common mode noise (suppressed by differential operation), IM2 injection circumvents gain and NF penalties.

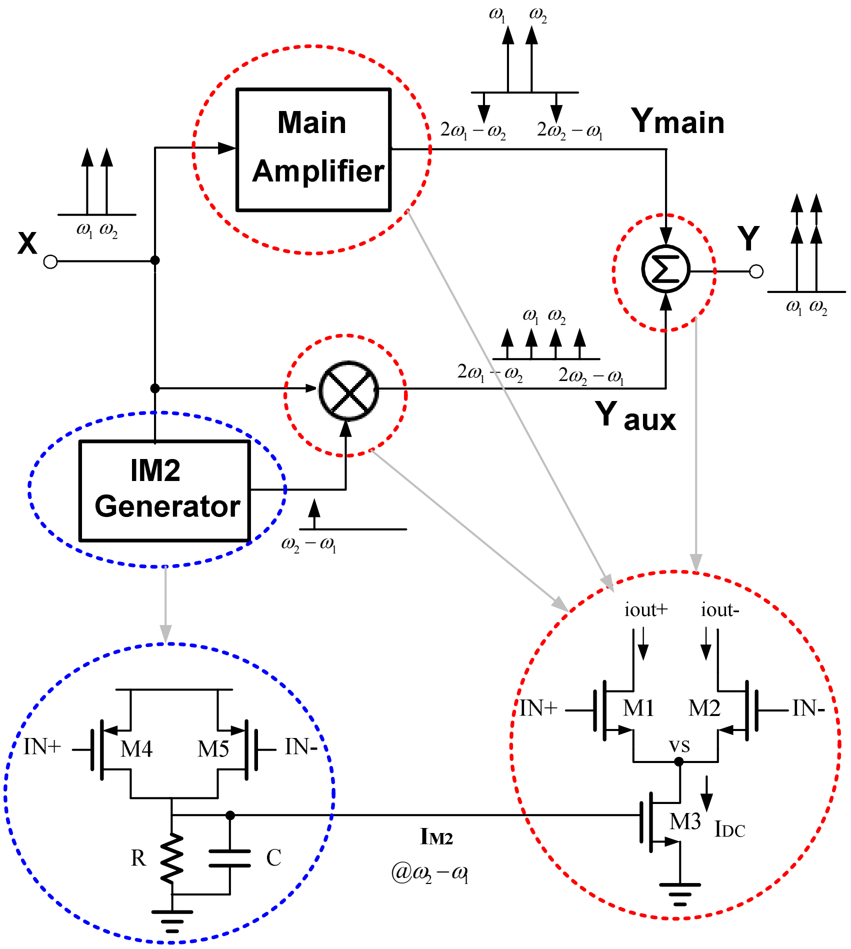


Fig. 2.16. Block diagram and basic cell implementation of “IM2 injection” [26].

Limitations of IM2 Injection include:

1) NMOS/PMOS transistors and resistors have independent PVT variations-hence more difficult to satisfy the IM3 cancellation criteria in (2.35) robustly.
2) Since R and C in the IM2 generator introduce extra phase shift, two tone

spacing must be smaller than the RC-filter cutoff frequency for negligible phase mismatch. Cancellation performance degrades as tone spacing increases.

3) Frequency components at $\omega_2 \pm \omega_1$ and $2\omega_{1,2}$ injected by the IM2 generator may fall into signal band and degrade the IIP2.
4) Noise from the IM2 generator is negligible only for differential LNAs, but would result in appreciable NF degradation for single-ended LNAs.

In short, IM2 injection applies chiefly to narrowband, differential systems with small two-tone spacing.

2.2.7 Noise/Distortion Cancellation

Noise/distortion cancellation parallels CG (M_A) and CS (M_B) stages, as shown in Fig. 2.17 [27]-[30]. The circuit is driven by a voltage at node "IN". The nonlinearity of M_A can be modeled as a current source between its drain and source controlled by both V_{gs} and V_{ds}. Hence, both the channel thermal noise and distortion of M_A flowing through the CG and CS paths are subtracted at the output, whereas the signal is added. It is required that the two paths through M_A and M_B are balanced for the noise/distortion current, i.e. $V_x = V_y$, we have:

$$g_{1,M_A} R_A = g_{1,M_B} R_B \text{ (differential output)} \tag{2.36a}$$

$$g_{1,M_{B1}} R_s = g_{1,M_{B2}} R_A \text{ (single-ended output)} \tag{2.36b}$$

Note that this technique can cancel all intrinsic distortion generated by M_A, including both g_m and g_{ds} nonlinearity, while previous techniques could only compensate g_m nonlinearity.

After cancelling the distortion from M_A, M_B's distortion dominates the residual nonlinearity, which comprises two terms: 1) M_B's intrinsic 3rd-order distortion and 2) 2nd-order interaction originating from the CG-CS cascade. Optimal biasing of M_B [28], [30], or employing complementary DS [29] could further improve the linearity.

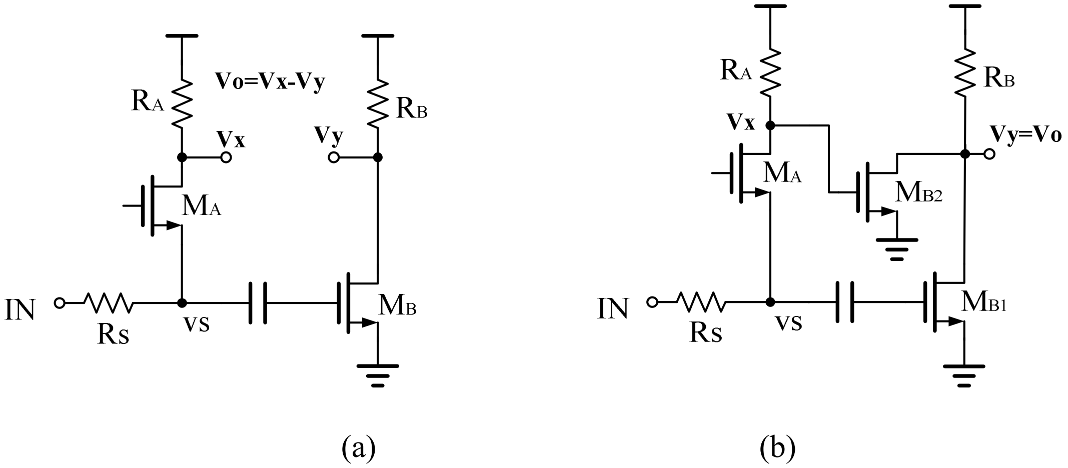


Fig. 2.17. Noise/distortion cancellation: (a) differential output [28], [30]; (b) single ended output [29].

2.2.8 Post-Distortion

Similar to the DS method, the Post-distortion (PD) technique also utilizes an auxiliary transistor's nonlinearity to cancel that of the main device, but it is more advanced in two aspects:

1) The auxiliary transistor is connected to the output of main device instead of directly to the input, minimizing the impact on input matching.
2) All transistors operate in saturation, resulting in more robust distortion cancellation.

Fig. 2.18 displays a conceptual diagram of PD as well as three implementations [31]-[33]. Note that the output impedance of i_{out} is not shown here for simplicity, but its effect is modeled. The auxiliary transistor M_B taps voltage v_2 and replicates the nonlinear drain current of the main transistor M_A, partially cancelling both 2nd- and 3rd-order distortion terms. The nonlinear drain currents of M_A and M_B can be modeled as:

$$i_A = g_{1A}v_1 + \underbrace{g_{2A}v_1^2 + g_{3A}v_1^3}_{f_{nonlin}(v_1)} \tag{2.37}$$

$$i_B = g_{1B}v_2 + g_{2B}v_2^2 + g_{3B}v_2^3 \tag{2.38}$$

Next, suppose v_2 is related to v_1 by:

$$v_2 = -b_1v_1 - b_2v_1^2 - b_3v_1^3 \tag{2.39}$$

where b_1-b_3 are generally frequency dependent and can be extracted from simulation. Note that (2.39) already models all the impedance at node v_2, including the output impedance of i_{out}, i_A, and i_B. In Fig. 2.18(a), the cascode devices were assumed to be ideal current buffers [31]. The two nonlinear currents i_A and i_B sum at node v_2, yielding i_{out}:

$$\begin{aligned} i_{out} &= i_A + i_B = \left(g_{1A} - b_1g_{1B}\right)v_1 \\ &+ \underbrace{\left(g_{2A} - b_1^2g_{2B} - b_2g_{1B}\right)v_1^2}_{\text{2nd-order distortion}} + \underbrace{\left(g_{3A} - b_1^3g_{3B} - g_{1B}b_3 - 2g_{2B}b_1b_2\right)v_1^3}_{\text{3rd-order distortion}} \end{aligned} \tag{2.40}$$

Note that in the PD method, both the main and auxiliary transistors operate in saturation with the same $g_{1,2,3}$ polarity. Hence, M_B partially cancels the linear term as well; however, it does not substantially degrade the gain/NF because M_B is designed to be more nonlinear than M_A (i.e. $g_{2,3B}/g_{1B} >> g_{2,3A}/ g_{1A}$). Finally, note that among the three implementations, Fig. 2.18 (b) and (d) might have better performance in practical implementations, since both M_A and M_B are NMOS, which can be matched very well in

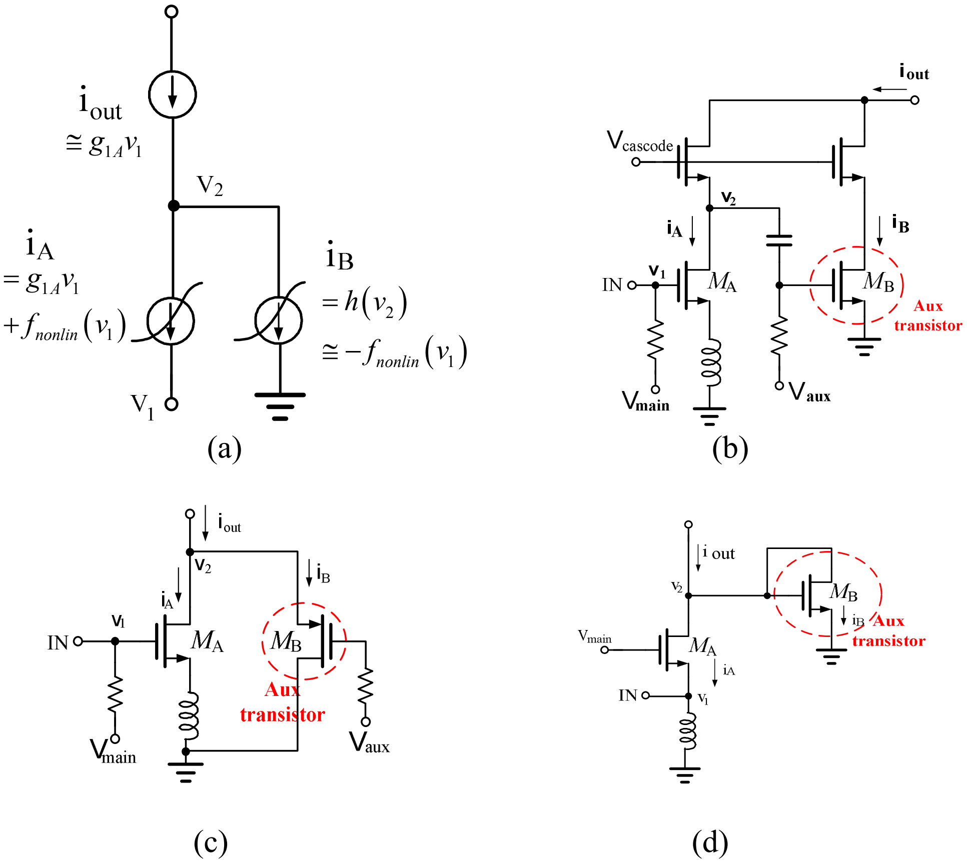


Fig. 2.18. Post-distortion: (a) conceptual view, (b) circuit implementation in [31]; (c) circuit implementation in [32]; (d) circuit implementation in [33].

layout. In Fig. 2.18(c), NMOS/PMOS transistors must have commensurate nonlinearity, but are hard to match across PVT. A detailed analysis and discussion of the topology in Fig. 2.18 (d) will be discussed later in Chapter IV.

2.2.9 Summary

Table 2.4 compares the IIP2/IIP3 improvement and gain/NF/power penalties of the previously discussed, state-of-the-art linearization techniques. We chose only one representative reference for each technique for brevity. The best performance per row has been marked with gray color. The modified DS method achieves the best IIP3 (>20dBm); the IM2 injection method yields minimum degradation in NF, gain, and power; and the PD method renders robust linearity improvement.

Note that transconductance linearization methods are inherently broadband, however, to apply it on wideband LNAs, we should match the delays and phases from the main and auxiliary paths, including input matching/loading network, at the desired frequency band, so that the distortion cancellation is carried out with sufficient accuracy. Most reported techniques (e.g. IM2 injection, modified DS, and harmonic termination) dealing with 2nd-order interaction are only limited to narrowband applications.

IIP2 calibration is another linearization technique that has been extensively reported for mixers, but still remains an open problem for LNAs. The concept of IIP2 calibration is to sense and correct the DC offset with an analog or digital feedback loop [34]-[37]. Some correction approaches for mixers include adjusting the LO bias [34], the load resistor/capacitor banks [35], the current source load [36], or injecting current at the

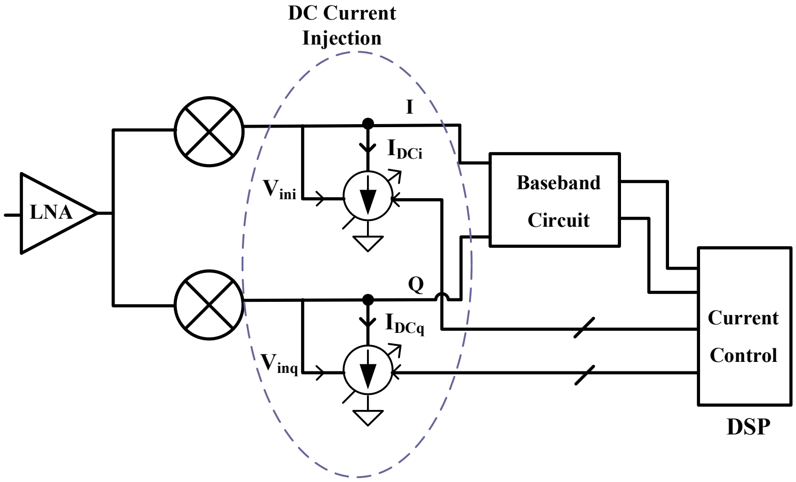
Table 2.4. Performance comparison of silicon-verified linearization techniques for CMOS LNAs

Linearization Technique	Harmonic Termination [15]	Optimum biasing [16]	Feedforward [17]	Derivative Superposition [19]	Modified DS [24]	Complementary DS ***[22]	IM2 Injection **[26]	Noise/Distortion Cancellation ***[30]	Post Distortion [33]
*IIP3/ΔIIP3	-4.4dBm/+2.5dB	+10.5dBm	5dBm/+13dB	2.7dBm/+13.4dB	2dBm/+20dB	3dBm	-10.4dBm/+10.6dB	>0dBm	5dBm/+9dB
*IIP2/ΔIIP2	N/A	N/A	N/A	N/A	N/A	+44dBm	N/A	>+20dBm	+10dBm/+10dB
*Gain/ΔGain	20.4dB/+2dB	14.6dB/0dB	18dB/-2.5dB	15.3dB/-0.4dB	16dB/-0.5dB	14dB	22dB/0dB	13-15.6dB	14.3dB/-1.7dB
*NF/ΔNF	1.92dB/0dB	1.8dB/0dB	2.6dB/+0.2dB	2.9dB/+0.1dB	1.4dB/+0.25dB	3dB	5.3dB/0dB	<3.5dB	2.7dB/+0.6dB
Power/ΔPower	16.2mW/0%	5.4mW/0%	22.5mW/+100%	20mW/+17.5%	23.4mW/+3.4%	34.8mW	19.6mW/+0.7%	14mW	2.6mW/+1%
Supply Voltage	1.8V	2.7V	3.0V	2.5V	2.6V	2.2V	1.5V	1.2V	1.3V
Frequency	2.2GHz	880MHz	900MHz	2.2GHz	900MHz	48-1200MHz	900MHz	0.2-5.2GHz	2.5-10GHz
Process	0.35μm	0.25μm	0.35μm	0.25μm	0.25μm	0.18μm	0.18μm	65nm	0.13μm
Robustness over PVT	moderate	poor	good	moderate	moderate	moderate	moderate	good	good
Wideband?	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes

* IIP3, IIP2, Gain, and NF value are the number before linearization; ΔIIP3, ΔIIP2, ΔGain, and ΔNF value are the improvement/degradation after linearization

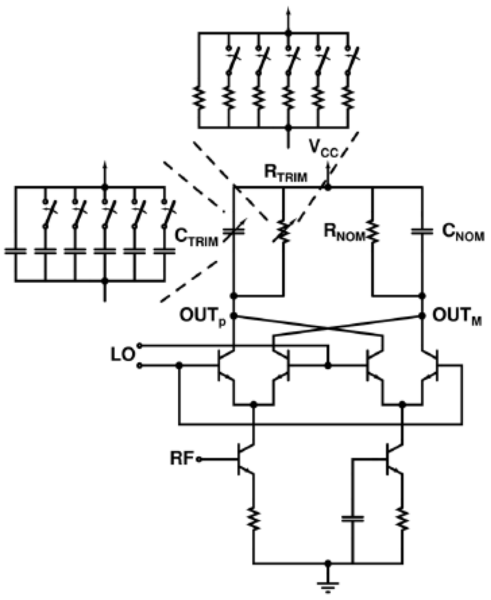
** Reported results are for LNA+Mixer

***only final results are reported, comparison results for with/without linearization circuitry are not available mixer output [37]. It might be possible to apply some of the methods currently employed in mixers to differential LNAs. Fig. 2.19(a) shows the concepts of “DC current injection”[37] It takes the down-converted blocker, and inject a dynamic dc offset at the mixer output, with an amplitude proportional to the blocker amplitude squared but in the opposite direction, thus effectively eliminates the 2nd-order component. Baseband ADCs measure the static and dynamic dc offset and determine the correct amount of injection. Several mismatch factors cause 2nd-order components in a Gilbert cell mixer, and the load resistor imbalance is one of them. Fig. 2.19(b) shows that tuning capacitors as well as resistors at the mixer output improves IIP2.



(a)

Fig. 2.19. Mixer IIP2 calibration schemes with (a) DC current injection (b) RC-calibration [35] (© 2004 IEEE).

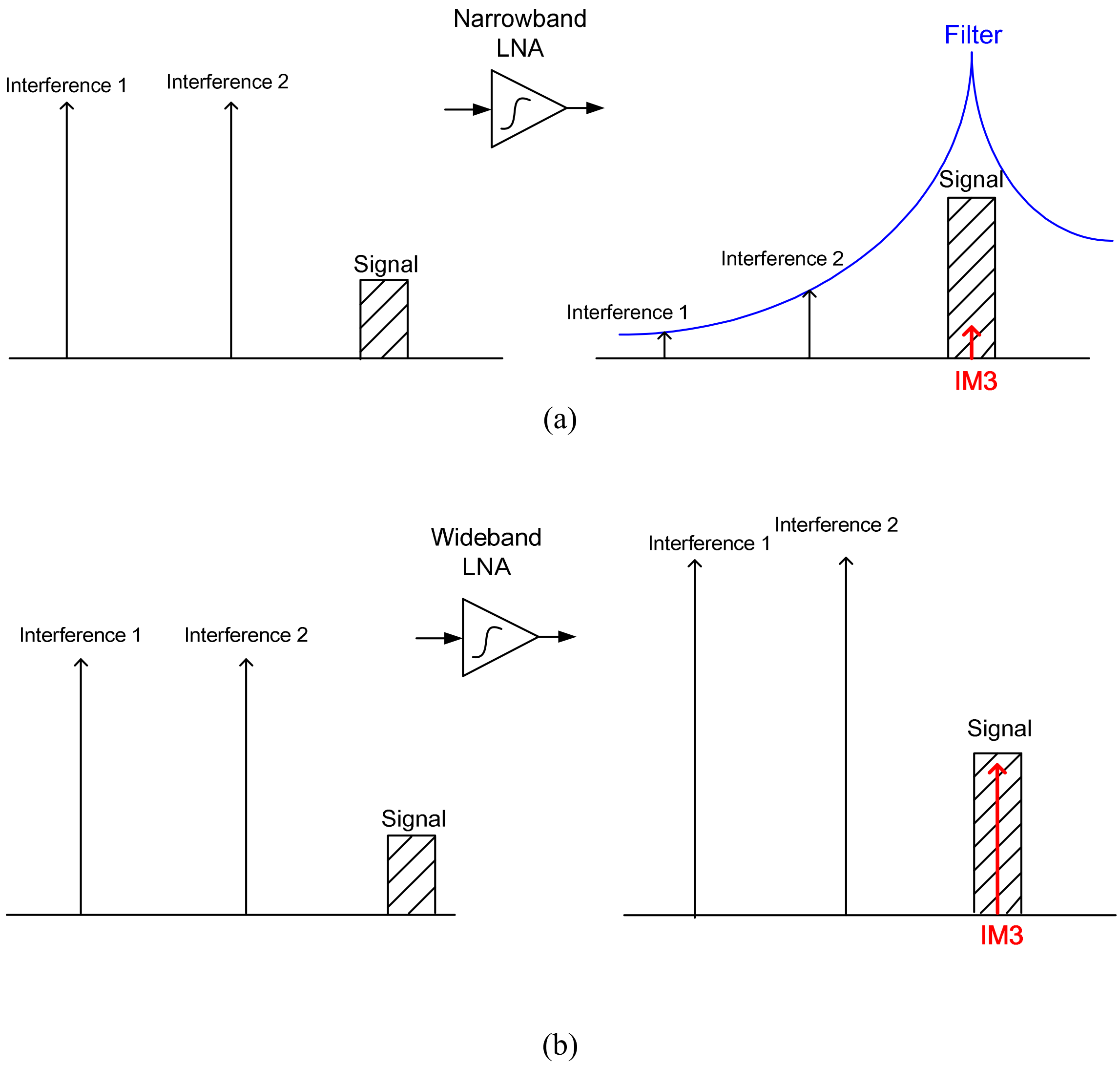


(b)

Figure 2.19 Continued.

2.3. New Issues for Wideband Applications

Growing research on reconfigurable multi-band/multi-standard and broadband transceivers has increased interest in broadband LNA design. In these transceivers, hundreds of channels could enter the LNA without any pre-filtering, acting as in-band interferers. As illustrated in Fig. 2.20, in narrowband receivers, the BPF suppresses interferences and preventing the LNA from generating a large IM3; however, for a multi-standard/wideband receiver, the BPF is broadband and interferences are not suppressed, creating a large distortion term on top of the main signal. If the LNA is not linear enough, the power of the distortion term may become comparable to that of the



(a)

(b)

Fig. 2.20. Distortion in (a) narrowband receivers (b) multi-standard/wideband receiver.

main signal, making it difficult to be recovered. Moreover, nearby radios and on-chip transmitter leakage cause increased adjacent blockers, creating severe cross-modulation, intermodulation, and desensitization. Therefore, a big design challenge for broadband LNAs is to achieve high linearity over a wide frequency range, lest the SNDR at the LNA output be dominated by distortion instead of noise. Furthermore, the old textbook argument that the LNA receives small input signal amplitude is not valid for wideband LNAs. We consider three main concerns: IIP2, P_{1dB}, and IIP2/IIP3 vs. two-tone frequency and spacing.

2.3.1 IIP2

Most linearization methods target narrowband applications and only cancel the 3rd-order distortion, since the 2nd-order nonlinearity is generally out of band in narrowband system. However, for wideband receivers, many channels are present concurrently and act as in-band interferences. Thus, the 2nd-order intermodulation products generated by certain combination of interferences are highly likely to fall into the signal band. Hence, broadband LNAs should have a good IIP2 as well as IIP3. Often, in applications like digital TV, the required IIP2/3 must be derived from a multi-tone test such as complex second-order distortion (CSO) and composite triple beat distortion (CTB) [22].

A fully differential LNA will improve IIP2, but requires a transformer, which is expensive for wideband systems. Other IIP2 improvement techniques include the complementary/differential DS method [22], [23], [29] and post-distortion [31]-[33]. Moreover, in deep submicron processes, biasing a CS-stage at the maximum gain yields a high IIP2 [30].

2.3.2 1dB Compression Point

The 1dB compression point (P_{1dB}) [7] , defined as the input signal level that causes the small-signal gain to drop by 1dB, quantifies the “large-signal” distortion of the circuit. P_{1dB} has traditionally not been a major concern for LNA designers because the

LNA typically has a small input signal. However, in wideband receivers, LNAs receive the accumulated power from multiple channels, which could range from -10 to 0dBm. For example, in the A/74 standard developed by the Advanced Television Systems Committee (ATSC), many transmitters are in close spectral proximity, so the receiver is exposed to more multicarrier adjacent energy. The maximum input power (the average of multiple tones) could even exceed 0dBm [42]. Furthermore, severe transmitter leakage, poor isolation between antennas, and single-tone blockers with large peak-to-average ratio all require a high signal-handling capability, i.e. high P_{1dB}, for the LNA to prevent desensitization, gain compression, and clipping.

IIP2/IIP3-improvement techniques typically only work over small signal ranges, and do not improve P_{1dB} because it is a large-signal parameter. At higher input amplitudes clipping occurs, and the P_{1dB} worsens due to limited supply voltage/DC-bias current.

P_{1dB}-improvement techniques include:

1) Increasing V_{dd} above nominal values to maximize the voltage headroom and performing substantial PVT simulation to guarantee breakdown/overstress will not occur.
2) Using low-f_T, thick-oxide transistors to handle larger voltage swings to allow even larger V_{dd}. Using such transistors degrades NF and high-frequency performance and raises cost.

Achieving high P_{1dB} with thin-oxide devices and low supply voltages remains an open problem. Some possible approaches include:

1) Cancel higher-order distortion, e.g. IM5 & IM7, since these become prominent at larger inputs and contribute to P_{1dB}.

2) Extend the effective input range of IM2/IM3 cancellation. One solution is to employ more auxiliary transistors in parallel in the DS method [21]. Note that weak-inversion transistors being turned on and off at large voltage swing will add more high-order harmonic components to the circuit. A more robust solution is to combine triode and weak inversion transistors as auxiliary transistors [21].

3) Add source degeneration at the cost of extra noise.

4) Dynamic bias/dynamic supply [43].

5) Reduce the output voltage swing to relax the limitation from nonlinear output conductance. One option is to use a low-impedance load for the LNA, for example by choosing a passive mixer over an active mixer as the following stage.]However, this choice requires a larger g_m stage and hence greater difficulty to linearize the transconductance.

2.3.3 IIP2/IIP3 vs. Two Tone Frequency and Spacing

Broadband LNAs have flat gain/NF over the whole bandwidth. Likewise, IIP2/IIP3 should also be relatively flat over the signal band. Therefore, while narrowband systems typically use a specific interference frequency and a small tone spacing for the two-tone test, broadband systems require IIP2/IIP3 to be examined at various two-tone-spacing and center frequencies [33]. Fig. 2.21 shows an example plot.

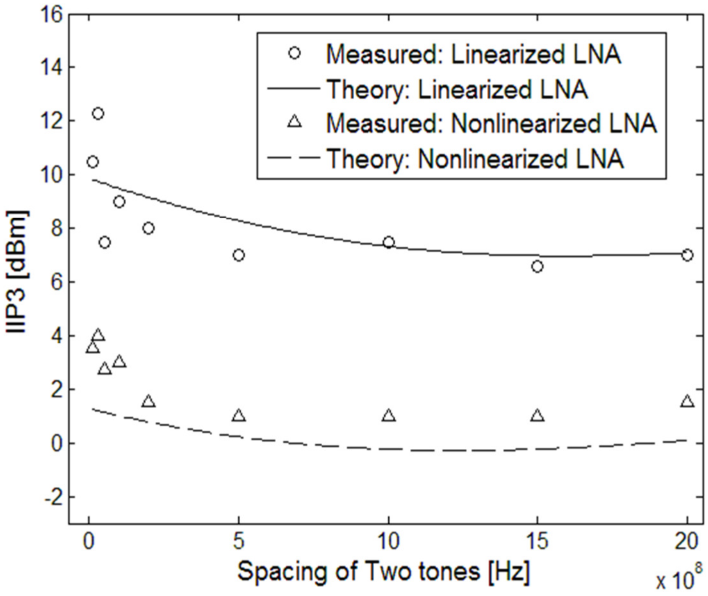


Fig. 2.21. Experimental and theoretical results of LNA IIP3 as a function of frequency spacing [33] (© 2009 IEEE).

Reactive components, such as those in the matching network, cause the frequency-dependence of IIP2/IIP3. Note that typically, this frequency-dependence is mild for operating frequencies below 1GHz, so it is more of a concern for UWB systems (3.1-10.6GHz) than for digital TV (54-880MHz), for example.

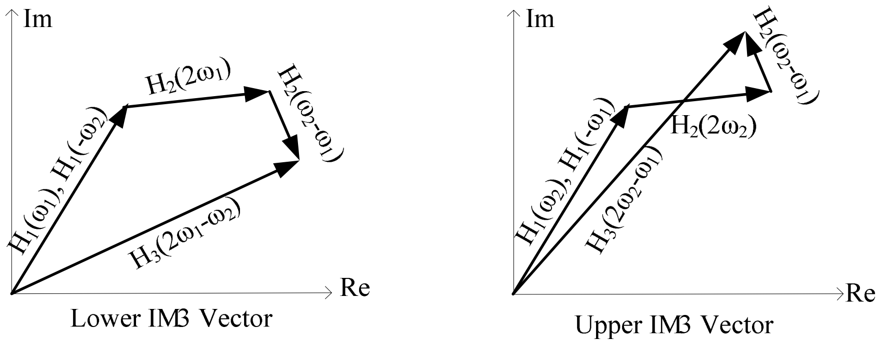
IIP2 depends on two-tone-spacing. For two input-signal tones at ω_1 & ω_2, the upper-frequency IM2 component is at $\omega_1+\omega_2$, while the lower-frequency component is at $\omega_1-\omega_2$. The IIP2 dependence on two-tone spacing is subtle when $\omega_1-\omega_2$ is very small. There are two situations for this dependence becomes more significant:

1) Large two-tone spacing, where larger frequency spacing yields correspondingly larger reactive effects.

2) Narrowband IM2 cancellation scheme. For example, in the complementary DS method with CG configuration shown in Fig. 2.11(b), the impedance from coupling capacitors increases with smaller two-tone spacing. Thus the AC-short condition worsens and degrades the IM2-cancellation effectiveness [22], [29].

The IIP3 dependence on two-tone spacing is mainly attributed to the "2nd-order interaction" as shown in (2.10). Therefore, the variations of $\Delta\omega$ cause the optimum point of the 2nd-order interaction cancellation to change, resulting in worse linearity. For example, in the IM2-injection method [26] (Fig. 2.16), the squaring circuit experiences more phase shift at larger two-tone spacing, which degrades IIP3. In the harmonic-termination method [9], IIP3 degrades noticeably at larger $\Delta\omega$.

Another major contributor to this IIP_3 dependence is the IM3 asymmetry, also called "sideband asymmetry". IM3 asymmetry is attributed to various types of memory effects [38]-[41], but for CMOS LNAs specifically, it is caused by the 2nd-order harmonic and difference frequency; i.e. the reactive part of the circuit impedance (e.g. termination impedance) at ω_2-ω_1 has a 180^o-out-of-phase contribution to the IM3 components at ($2\omega_1$-ω_2) and ($2\omega_2$–ω_1). This concept is qualitatively illustrated by the vector diagram in Fig. 2.22 [39], where the $H_{1,2,3}$ refers to the 1st-, 2nd-, and 3rd-order Volterra-Series coefficients. The IM3 components at ($2\omega_2$-ω_1) and ($2\omega_1$-ω_2) have different imaginary parts (i.e. reactance), resulting in IM3 asymmetry.



Lower IM3 Vector

Upper IM3 Vector

Fig. 2.22. Vector diagram showing the 180° out-of-phase contribution of ω_2-ω_1 term on the upper and lower IM3 components [39].

Note that this IM3 asymmetry depends on bias and frequency. For very small two-tone spacing, it is hard to see any IM3 asymmetry since the reactive-impedance effect at $\Delta\omega$ is negligible; but for larger $\Delta\omega$, the reactive impedances at the 2nd-harmonic frequency also contribute differently to the lower/upper IM3 components, which worsens the IM3 asymmetry [9] and also indicates a more obvious IIP3 dependence on two-tone-spacing. Choosing a proper bias, and minimizing the "2nd-order interaction" can help to reduce this IM3 asymmetry [41]. Note that in the multi-tone case, adjacent channel power ratio (ACPR) asymmetry is defined correspondingly.

2.4 LNA Linearization in Deep Submicron Technology

2.4.1 Nonlinearity from Output Conductance g_{ds}

Distortion of MOS transistors is mainly caused by the nonlinear transconductance (g_m) and output conductance (g_{ds}). Previously published linearization techniques mainly focus on linearizing g_m, assuming that: 1) drain current i_{ds} is controlled only by the gate-source voltage V_{gs}; 2) g_{ds} nonlinearity is negligible. These assumptions are valid for small load resistance, small voltage gain, small input signal, and a drain-source voltage (V_{ds}) sufficiently large that the small-signal variation of V_{ds} does not appreciably perturb the bias point.

However, as technology scales down, the g_{ds} nonlinearity becomes more prominent. Current i_{ds} is controlled not only by V_{gs} but also the V_{ds}, which can be approximated by the two-dimensional Taylor series [6], [30]:

$$\begin{aligned}i_{ds}\left(V_{gs},V_{ds}\right)&=g_1V_{gs}+g_2V_{gs}^2+g_3V_{gs}^3+g_{ds1}V_{ds}+g_{ds2}V_{ds}^2+g_{ds3}V_{ds}^3\\&+c_{(1,1)}V_{gs}V_{ds}+c_{(2,1)}V_{gs}^2V_{ds}+c_{(1,2)}V_{gs}V_{ds}^2\end{aligned}\tag{2.41}$$

where g_i is the i^{th}-order transconductance as defined in (2.7); g_{dsi} represents the nonlinear output conductance effect which is proportional to the i_{ds} derivatives with respect to V_{ds}; $c_{(m,n)}$ is the cross-modulation term describing the dependence of g_i on V_{ds} or g_{dsi} on V_{gs}, as formulated in (2.42):

$$g_{dsi}=\frac{1}{i!}\frac{\partial^i I_{DS}}{\partial V_{DS}^i}\ ,\qquad c_{(m,n)}=\frac{1}{m!n!}\frac{\partial^{m+n}I_{DS}}{\partial V_{GS}^m\partial V_{DS}^n}\tag{2.42}$$

To characterize the g_{ds} nonlinearity for a single transistor, we fix its V_{gs} at 0.5V and inversion level i_f as 30, and sweep the V_{ds}, by taking the first three derivatives of the drain-source DC current i_{ds} with respect to V_{ds} (as defined in equation (2.42)) at every DC bias point, we can obtain Fig. 2.23. It is observed that the drain current is modulated a lot by V_{ds}. g_{ds3} is large when the transistor operates at small V_{ds}; while it decreases for large V_{ds} values. Design hints for minimizing the gds-induced nonlinearity are discussed in section 2.5.2.

Here we assume a negligible nonlinearity contribution from g_{mb}, otherwise three dimensional Taylor series should be used instead. From (2.41), the distortion is contributed by four parts:

1) g_m nonlinearity due to nonlinear i_{ds}-V_{gs} relation.

2) g_{ds} nonlinearity from channel length modulation effect. Note that g_{ds} contributes less nonlinearity when device operates deeper into saturation region.

3) the dependence of g_m on V_{ds}, (partially due to the drain induced barrier lowering (DIBL) effect [30].

4) the dependence of g_{ds} on V_{gs}, especially in saturation region [6].

The cross modulation effect remains fairly constant for a broad range of V_{gs}, while g_m is more linear and g_{ds} becomes more nonlinear as V_{gs} increases, V_{ds} decreases, and transistors operate close to the linear region. [30] demonstrated that by choosing a proper V_{gs} for a CS stage, the $V_{gs}.V_{ds}$ cross-term $c_{1,2}$ cancels the intrinsic 2^{nd}-order distortion (g_2) , resulting in a high IIP2. Note that when g_{ds} nonlinearity dominates (i.e. output limited), the tradeoff between gain and linearity becomes more severe.

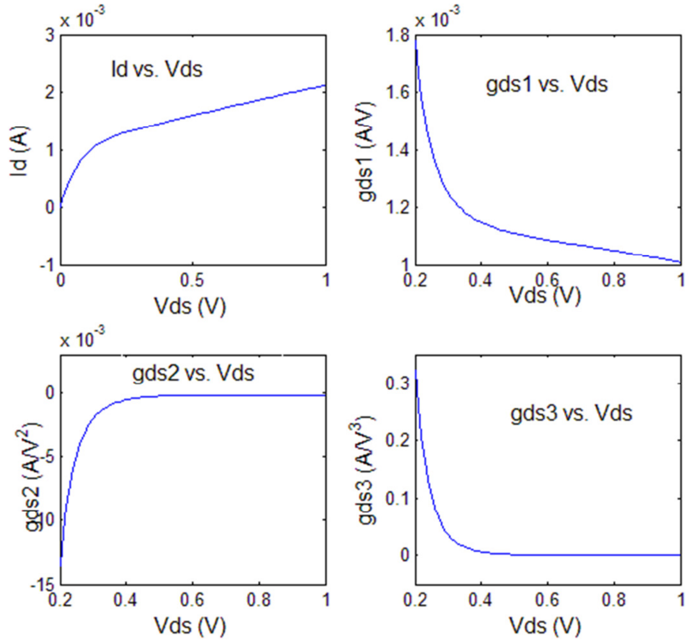


Fig. 2.23. NMOS output conductance nonlinearity characteristics (UMC 90nm CMOS process, W/L = 20/0.08μm, $V_{gs} = 0.5V$, $V_{th} = 0.26V$).

2.4.2 MOSFET Capacitance

For the most part, the capacitances of a saturation-region transistor are linear at an operating frequency less than $f_T/10$ [6]. Therefore they do not directly contribute to distortion [44]. However, if a strong blocker is present (e.g. in the order of 0dBm), the input capacitance C_{gg} varies significantly around the threshold voltage, and its nonlinearity becomes significant. The expression for C_{gg} is as follows [49]:

$$C_{gg} = \frac{2}{3} C_{ox} \left[1 + \exp\left(-\frac{V_{GS} - V_T}{m\phi_T} \right) \right]^{-1} \tag{2.43}$$

where m determines the sub-threshold slope (m=1.3). Also, as previously mentioned, the gate-drain capacitance (C_{gd}) provides a feedback path for the "2nd-order interaction," and this C_{gd} effect becomes more visible as the load impedance increases. At high frequency,

C_{gd} and the drain-bulk capacitor (C_{db}) reduce total output impedance and hence the output voltage swing, helping to mitigate the nonlinear g_{ds} effect. Therefore, g_m nonlinearity dominates at high frequency, while g_{ds} nonlinearity dominates at low frequency [44]. However, in those circuits where capacitive components are tuned out for a matched load, g_{ds} nonlinearity is still prominent at high frequencies. The substrate affects linearity through C_{db} with higher operating frequency, and this effect varies with different substrate-doping profiles [44]. Generally, IIP3 improves as substrate doping increases [47]. The effect from substrate leakage current can typically be neglected [45].

Table 2.5. Dominant contributor to distortion under various conditions

	g_m	g_{ds}
Small load resistance	√	
Large load resistance		√
Small voltage gain Av	√	
Large voltage gain Av		√
High frequency	√	
Low frequency		√
Saturation region	√	

Table 2.5 provides a summary: the mark “√” denotes whether g_m or g_{ds} dominates the distortion under the given conditions.

2.4.3 Impact of Technology Scaling on Linearity

As channel length decreases, the velocity saturation effect becomes prominent, i.e. the drain current saturates at smaller V_{ds}. Thus, the long-channel equation for drain current in saturation region needs to be modified as [6]:

$$I_{ds} = \frac{\mu C_{ox}}{2}\frac{W}{L}\left(V_{gs} - V_t\right)\left[\left(V_{gs} - V_t\right) \square \left(LE_{sat}\right)\right]$$
$$\approx \frac{\mu C_{ox}}{2} W\left(V_{gs} - V_t\right)E_{sat} \quad \text{(for small L)} \tag{2.44}$$

where E_{sat} is the field strength at which the carrier velocity drops to half the value extrapolated from low-field mobility. g_m becomes more linear:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{\mu C_{ox}}{2} W E_{sat} \tag{2.45}$$

The vertical-field mobility degradation effect also helps to linearize g_m in DSM process. The long-channel equation for drain current can be modified as:

$$I_{ds} = \frac{\mu C_{ox}}{2}\frac{W}{L}\frac{\left(V_{gs} - V_t\right)^2}{1 + \theta\left(V_{gs} - V_t\right)} \tag{2.46}$$

where $\theta \propto 1/t_{ox}$ models vertical-field mobility degradation. (2.46) reduces to $I_{ds} \propto \left(V_{gs} - V_t\right)$ as $V_{gs} - V_t$ increases, resulting in a linear I-V curve, and g_m becomes constant with respect to bias voltage:

$$g_m \approx \frac{\mu C_{ox}}{2}\frac{W}{L}\frac{1}{\theta} \tag{2.47}$$

On the other hand, g_{ds} is more nonlinear for shorter channel length, as proven by the experimental data in [44], and it can be expressed as follows [50] :

$$g_{ds} = 2I_{DS}\left(1-\frac{V_{GS}}{V_T+\Delta V_T+\gamma V_{DS}}\right)\cdot\frac{\gamma V_{GS}}{\left(V_T+\Delta V_T+\gamma V_{DS}\right)^2}\cdot\tanh\left(\alpha V_{DS}\right)\cdot\left(1+\lambda V_{DS}+\beta V_{GS}\right)$$
$$+\alpha I_{DS}\left(1-\frac{V_{GS}}{V_T+\Delta V_T+\gamma V_{DS}}\right)^2\left[1-\tanh^2\left(\alpha V_{DS}\right)\right]\cdot\left(1+\lambda V_{DS}+\beta V_{GS}\right)$$
$$+\lambda I_{DS}\left(1-\frac{V_{GS}}{V_T+\Delta V_T+\gamma V_{DS}}\right)^2\cdot\tanh\left(\alpha V_{DS}\right)$$

(2.48)

where α determines the drain voltage where the drain current characteristic saturates, γ simulates the effective threshold voltage displacement as a function of V_{DS}, and ΔV_T is the geometric shift in threshold voltage. Furthermore, the reduced supply/V_{dsat} values result in the device being biased closer to the triode-saturation boundary, which worsens the g_{ds} nonlinearity. Consequently, the maximum OIP3 occurs with smaller load impedance (which mitigates the distortion contribution from nonlinear g_{ds}) and the peak IIP3 shifts to lower V_{gs} [47], since a smaller overdrive voltage allows the device to stay far away from the triode-saturation boundary while still keeping g_{ds} nonlinearity small.

The “sweet spot” in the optimal biasing technique(discussed in Section 2.2.3) will systematically shift to higher bias-current density I_{ds}/W (i.e., larger overdrive voltage) as technology scales down [6], which means larger power is required to preserve linearity.

As oxide thickness decreases, poly-gate depletion increases, and the nonlinear gate capacitance develops strong 2nd-order derivatives (C_{g2}) with respect to V_{gs}, which contribute to significant 3rd-order distortion (g_3) in drain current, as shown below [46]:

$$I_{ds} = Qv \approx Qv_{sat} \tag{2.49}$$

$$g_3 = \frac{\partial^3 I_{ds}}{\partial V_{gs}^3} \approx \frac{\partial^3 Q}{\partial V_{gs}^3} v_{sat} \approx \frac{\partial^2 C\left(V_{gs}\right)}{\partial V_{gs}^2} v_{sat} = C_{g2} v_{sat} \quad (2.50)$$

where Q is the channel charge density along the current direction, v is the carriers' velocity, and v_{sat} is the saturated velocity for sufficiently high field. Thus, distortion increases with thinner oxides.

DIBL becomes more severe in deep submicron process, besides a V_{ds}-dependent g_m, DIBL also affects the linearity by changing the effective V_{th}[48] . Measured results in [48] shows that the distortion is more sensitive to DIBL effect when the drain voltage increases and the MOSFET operates in moderate region (i.e. V_{gs} is slightly higher than V_{th}).

Finally, each process has a "low frequency limit" (LFL), below which the MOSFET exhibits fairly frequency-independent linearity. LFL is closely related to the device speed and can be approximated as $f_T/5$[48]. Therefore, it is easier to achieve IIP2/IIP3 flatness over the signal band in smaller-size technology.

In summary, as technology scales down, the transistor intrinsic gain g_m/g_{ds} decreases; lower supply voltages reduce the headroom and can lead to greater nonlinearity from g_{ds}, necessitating multidimensional Taylor analysis to model the nonlinear I_{ds}. Higher-order effects such as DIBL, velocity saturation, and poly-gate depletion all affect linearity. A key challenge resides in delivering high linearity with core transistors and with a low supply voltage in the DSM processes.

2.5 Remarks for High Linearity LNA Design

Besides applying explicit linearization techniques to the circuit, some general guidelines are helpful for designing a high-linearity LNA.

2.5.1 To Reduce g_m-induced Distortion

From (2.46), the low-frequency expressions for second- and third-order intercept points A_{IIP2} and A_{IIP3} are [7]:

$$A_{IIP2}^2 = \left|\frac{g_1}{g_2}\right| = \left|V_{dsat}\left(2+\theta V_{dsat}\right)\left(1+\theta V_{dsat}\right)\right| \qquad (2.51)$$

$$A_{IIP3}^2 = \frac{4}{3}\left|\frac{g_1}{g_3}\right| = \left|\frac{4}{3}\frac{V_{dsat}}{\theta}\left(2+\theta V_{dsat}\right)\left(1+\theta V_{dsat}\right)^2\right| \qquad (2.52)$$

where $V_{dsat} = V_{gs} - V_{th}$. Equations (2.51) and (2.52) indicate that increasing V_{dsat} improves both IIP2 and IIP3. Therefore, given sufficient voltage headroom, maximizing V_{dsat} and minimizing transistor sizes (W/L) helps to minimize parasitics and to linearize the circuit at the cost of increased DC current.

2.5.2 To Reduce g_{ds}-induced Distortion

As discussed in section 2.4.1, g_{ds} becomes more nonlinear as V_{ds} decreases and transistor operates towards the linear region; therefore, increasing supply voltage mitigates the g_{ds} effect, allows larger output swing and hence improves P_{1dB}. But the voltage drop across core transistors must be ensured not to exceed the safe operation value.

Provided sufficient voltage headroom, adding cascode device allows the output impedance from transistors to be much larger than load resistor, yielding a more linear output load.

With cascode transistor, most of the output swing will show as V_{ds} variation at the cascode transistor, while the V_{ds} of input transistor remains relatively constant. Therefore, the nonlinear output conductance of the cascode transistor has more contribution to the overall distortion. It is helpful to bias the cascode transistor at smaller V_{gs} (i.e. lower overdrive voltage) to tolerate a larger swing at the drain.

If supply voltage cannot be increased, we can:

1) use longer channel length to reduce the channel length modulation effect (assuming speed is not an issue);
2) reduce the load resistance of the LNA, which may affect the design of other building blocks in the receiver.

2.5.3 To Reduce Second-order Distortion

1) Biasing a CS-stage at the maximum gain yields a high IIP2 in deep submicron process [30].

3) Biasing the device for maximum f_T yields minima in the 2^{nd}-harmonic [48]; this intrinsic distortion cancellation results from opposite contributions of gate capacitance and g_m, as the device enters the linear region from saturation.

2.5.4 Other Tips

For inductively degenerated CS-LNAs, we can reduce Q to mitigate the "Q boosting" effect [7], provided that there is enough margin in NF and gain. Since C_{gs} has negligible effect on linearity, an external capacitor can be added in parallel with C_{gs} to allow more freedom for input transistor sizing. On the other hand, CG-LNAs generally provide better linearity than CS-LNAs [33] because it doesn't have this "Q boosting".

Use cascode transistors whenever possible because they:

1) reduce 2nd-order interaction through C_{gd}

2) reduce the voltage swing across each active device, improving reliability for DSM devices.

2.6 Conclusions

We have reviewed eight categories of CMOS LNA-linearization techniques and discussed the tradeoffs among linearity, power, and PVT variations. We subsequently discussed wideband-LNA-linearization issues for the emerging broadband transceivers, noting that IIP2 is becoming just as important as IIP3, and that improving P_{1dB} is also necessary for wideband applications to improve high-signal-handling capability. Issues in deep submicron processes, such as nonlinear output conductance were examined. A key challenge resides in delivering high linearity with core transistors and low supply voltage in the deep submicron processes. Linearization techniques for cancelling higher-order distortion terms (beyond 3rd order), linearizing output conductance, and improving LNA P_{1dB} still remain open problems. Finally, we presented general design guidelines for high-linearity LNAs.

CHAPTER III

PROPOSED LINEARIZATION TECHNIQUE FOR A DIFFERENTIAL CASCODE LNA*

3.1 Introduction

Due to the low cost and easy integration, CMOS is widely used to design wireless systems especially in the radio frequency region. The Low Noise Amplifier (LNA) serves as the first building block of the wireless receiver. It needs to amplify the incoming wireless signal without adding much noise and distortion. The noise performance of the LNA dramatically influences the overall system noise performance. The inductively degenerated CS-LNA [7], [51] is widely used due to its superior noise performance. A common gate LNA (CG-LNA) can easily achieve the input impedance matching, but suffers from poor noise performance [54]. The capacitive cross-coupling technique for CG-LNA [55]-[57] partially cancels the noise contribution of the common gate transistor at the output, which improves the noise performance of the CG-LNA. On the other hand, due to the existence of the parasitic capacitance at the source of the cascode transistor, the cascode transistor's noise influences the overall noise performance of the CS-LNA [58]-[62]. In [61], a layout technique to merge the main transistor and the cascode transistor can reduce the cascode transistor noise contribution.

*©[2008] IEEE. Reprinted, with permission, from “A Noise Reduction and Linearity Improvement Technique for a Differential Cascode LNA”, by Xiaohua Fan, Heng Zhang, and Edgar Sánchez-Sinencio, *IEEE Journal of Solid-State Circuits,* vol. 43, no. 3. pp. 588-599, March 2008.

Additional inductors can be added at the drain of the main transistor to cancel the effect of the parasitic capacitance, thus improving the noise performance of the LNA [58]-[54] at the cost of larger area for the on-chip inductors.

In this chapter, a noise reduction inductor combined with the capacitive cross-coupling technique is proposed to improve the noise and linearity performance of the differential cascode LNA. It can reduce the noise and nonlinearity contributions of the cascode transistors with a smaller inductor compared with the typical inductor based technique [58]-[60]. The capacitive cross-coupling technique used in the cascode transistors increases the effective transconductance of the cascode transistors, further improves the linearity of the LNA, and also reduces the Miller effect of the gate drain capacitance of the main transistor.

Section 2.2 describes the basic inductively degenerated CS-LNA, analyzes the noise influence of the cascode transistors, and shows the conventional inductor based noise improvement technique. Section 2.3 discusses the original capacitive cross-coupling technique [55]-[57] for CG-LNA, and proposes its application combined with inductor in the cascode transistors of the differential cascode CS-LNA. It also gives the theoretical foundations of the LNA noise reduction with the proposed technique. Section 2.4 discusses the LNA linearity improvement with the proposed technique. Section 2.5 addresses the effects of the proposed technique on the LNA S11 and gain. The measurement results are presented in section 2.6 and section 2.7 provides conclusions.

3.2 Background and Previous Work

The LNA noise performance dominates the overall noise performance of the receiver. The inductively degenerated CS-LNA is widely used due to its superior noise performance.

3.2.1 Inductively Degenerated CS-LNA

The typical inductively degenerated CS-LNA is shown in Fig. 3.1, where all parasitic capacitances other than the gate-source capacitances of M_1 and M_2 are ignored for simplicity. It uses an inductor L_s to generate the real impedance to match the input impedance to 50Ω, which results in good noise performance [7], [51]-[52]. If the resistive losses in the signal path, the gate resistance, and the parasitic capacitances except gate-source capacitances are ignored, the overall input impedance of CS-LNA can be simplified to (3.1), where g_{m1} is the transconductance of M_1.

$$Z_{in}(s) \approx sL_g + sL_s + \frac{1}{sC_{gs1}} + g_{m1}\frac{L_s}{C_{gs1}} \tag{3.1}$$

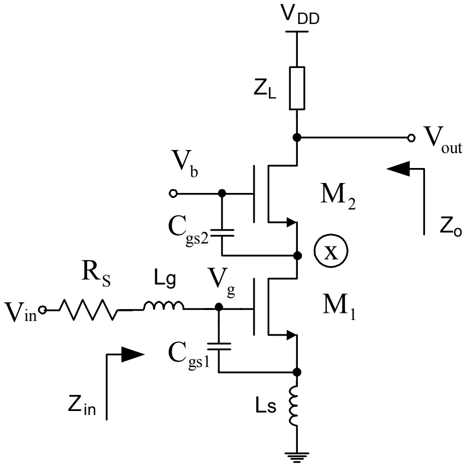


Fig. 3.1. Inductively degenerated cascode CS-LNA.

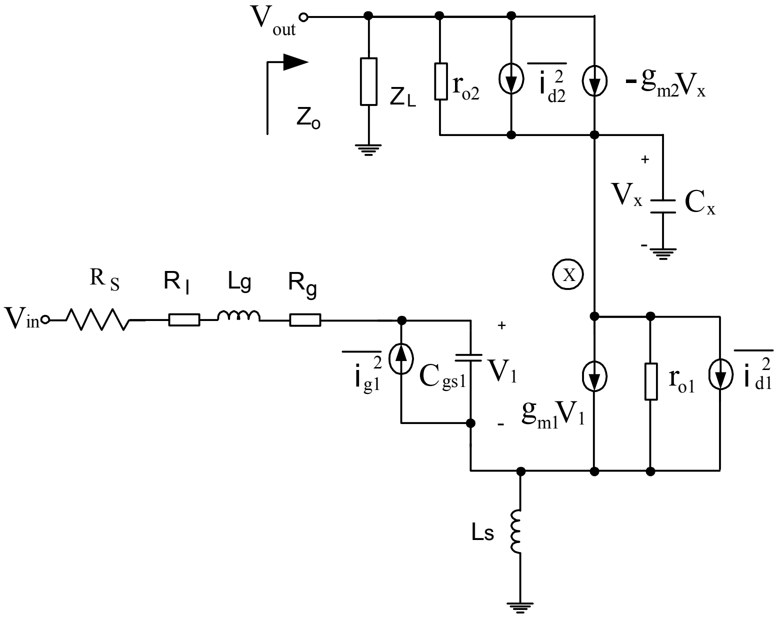


Fig. 3.2. Small signal model of cascode CS-LNA for noise analysis.

The small signal model of the inductively degenerated cascode CS-LNA is shown in Fig. 3.2, where C_{gd} and g_{mb} of the transistors are ignored for simplicity. The capacitor C_x represents all the parasitic capacitances at node X. It is estimated as:

$$C_x \approx C_{gs2} + C_{sb2} + C_{db1} \tag{3.2}$$

If the noise contribution from the cascode stage is ignored, the noise factor F_1 of the cascode CS-LNA becomes [7], [51]-[52]:

$$F_1 = 1 + \frac{R_1}{R_s} + \frac{R_g}{R_s} + \frac{\gamma}{\alpha}\frac{\chi}{Q_L}\frac{\omega_o}{\omega_T} \tag{3.3}$$

$$\chi = 1 - 2|c|\sqrt{\frac{\delta\alpha^2}{5\gamma}} + \frac{\delta\alpha^2}{5\gamma}(1 + Q_L{}^2) \tag{3.4}$$

$$Q_L = \frac{\omega_o(L_s + L_g)}{R_s} = \frac{1}{\omega_o C_{gs1} R_s} \tag{3.5}$$

$$c = \frac{\overline{i_{ng} i_{nd}^*}}{\sqrt{\overline{i_{ng}^2}\,\overline{i_{nd}^2}}} \approx -0.395j \tag{3.6}$$

where R_s is the input voltage source resistance, R_1 represents the series resistance of the inductor L_g, R_g is the gate resistance of M_1, ω_0 is the operating frequency, c is the correlation coefficient between the gate noise i_{ng} and the thermal noise i_{nd}, and α, γ, and δ are bias-dependant parameters [7], [51]-[52]. The existence of the parasitic capacitance C_x reduces the gain of the first stage, which makes the noise contribution from the cascode stage (Fc) larger. Thus, the noise factor of the cascode CS-LNA [59]

$$F = F_1 + F_c \approx F_1 + 4R_s\gamma_2 g_{do2}\left(\frac{\omega_o{}^2 C_x}{\omega_T g_{m2}}\right)^2 \tag{3.7}$$

where $\omega_T = g_{m1} / C_{gs1}$, g_{do2} is the zero-bias drain conductance of M_2 and γ_2 is the bias-dependent factor. Same as in [59], the noise sources of the first stage include the gate induced noise and drain noise sources, but only the drain noise of the second stage is modeled [55]-[59]. From (3.2) and (3.7), it can be observed that C_x increases the noise factor of the LNA. The exact noise contribution from M_2 varies in different designs, for this design in 0.35μm CMOS process, it adds 0.5dB to the overall 2.5dB LNA NF.

3.2.2 Existing Solution to Reduce Noise

The parasitic capacitance C_x can be reduced by merging the main transistor and the cascode transistor in the layout [61]. In [58]-[54], an additional inductor L_{add} was added to cancel the effect of C_x at the frequency of interest. As a result, if the r_{o1} and r_{o2} of M_1 and M_2 are large enough, the noise current generated by the cascode transistor M_2 adds negligible noise current to the output.

The large area requirement of on-chip inductor is a big concern for on-chip integration. For a typical 0.35μm CMOS technology, the parasitic capacitance for a 200μm/0.4μm NMOS transistor is nearly 0.3pF. Thus, it requires an inductor around 14nH to resonate at 2 GHz. In the advanced CMOS technology, it requires even larger inductor values. In addition, the poor quality factor of the on-chip inductor increases the overall noise figure of the LNA.

In this chapter, we propose a technique to significantly reduce the noise and nonlinearity contribution of the cascode transistors as well as the value of L_{add}.

3.3 LNA Noise Reduction with the Proposed Technique

The CG-LNA can achieve wideband input impedance matching, but suffers from poor noise performance. To alleviate this problem, a capacitive cross-coupling technique was proposed in [55]-[57] for CG-LNA. It can boost the transistor transconductance with passive capacitors, as shown in Fig. 3.3. If the gate-bulk and gate-drain capacitances are ignored, the effective transconductance and input capacitance of the LNA are here derived as (3.8) and (3.9). When $C_c >> C_{gs}$, the effective transconductance is doubled, and the input capacitance is increased by four times.

$$G_{m,eff} = \frac{2C_c}{C_{gs} + C_c} g_{m1} \tag{3.8}$$

$$C_{in} = \frac{4C_c}{C_{gs} + C_c} C_{gs} = 2\frac{G_{m,eff}}{g_{m1}} C_{gs} \tag{3.9}$$

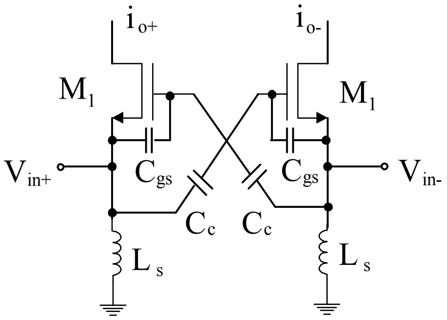


Fig. 3.3. A capacitor cross-coupled differential CG-LNA.

The inductively degenerated cascode CS-LNA can be considered as a CS-CG two stage LNA. The CS stage is designed to achieve the input impedance matching and also to obtain best noise performance. The input voltage signal is converted to current through the CS transistor. The cascode transistor works as a CG stage. It is designed mainly to reduce the Miller effect of the parasitic gate-drain overlap capacitance in the CS transistor. It also helps to increase the output impedance and to improve the input-output isolation.

An additional inductor L_{add} combined with the capacitive cross-coupling technique is applied to the cascode transistors of the differential LNA to reduce the noise. The proposed topology is implemented in a fully-differential inductively degenerated CS-LNA as shown in Fig. 3.4. The input admittance at node X is given by $G'_{m,eff}(j\omega) + jB'_{eff}(j\omega)$, where the effective transconductance of the cascode transistor is expressed as:

$$G'_{m,eff} = \frac{2\omega C_c - \dfrac{1}{\omega L_{add}}}{\omega C_c + \omega C_{gs2} - \dfrac{1}{\omega L_{add}}} g_{m2} \tag{3.10}$$

Note that in our proposed approach, the cross-coupled capacitors are applied to the cascode transistors. The equivalent input susceptance at node X is not purely capacitive, which can be derived as:

$$B'_{eff} \cong \frac{4\omega C_c}{\omega C_c + \omega C_{gs2} - \dfrac{1}{\omega L_{add}}} \omega C_{gs2} + \frac{(\omega C_c + \omega C_{gs2}) \times (-\dfrac{1}{\omega L_{add}})}{\omega C_c + \omega C_{gs2} - \dfrac{1}{\omega L_{add}}} + \omega C_{db1} + \omega C_{gd1} + \omega C_{sb2} \tag{3.11}$$

where other parasitic capacitances are ignored. From (3.10), if $\omega C_c >> \omega C_{gs2}$ and $\omega C_c >> \omega C_{gs2} - \frac{1}{\omega L_{add}}$, the effective transconductance is doubled and the equivalent susceptance from (3.11) becomes

$$B'_{eff} \cong \frac{\omega C_c (4\omega C_{gs2} - \frac{1}{\omega L_{add}})}{\omega C_c + \omega C_{gs2} - \frac{1}{\omega L_{add}}} + \omega C_{sb2} + \omega C_{gd1} + \omega C_{db1} \qquad (3.12)$$

At $\omega = \omega_o$, when (3.12) equals to zero, the capacitive effect at node X is mainly eliminated, which leads to

$$L_{add} \approx \frac{1}{\omega_o{}^2 (4C_{gs2} + C_{sb2} + C_{gd1} + C_{db1})} \qquad (3.13)$$

Using the small signal model, the noise figure of the cascode LNA yields

$$F' = F_1 + F'_c = F_1 + 4R_s \gamma_2 g_{do2} \left(\frac{\omega_o B'_{eff}}{\omega_T G'_{m,eff}} \right)^2 \qquad (3.14)$$

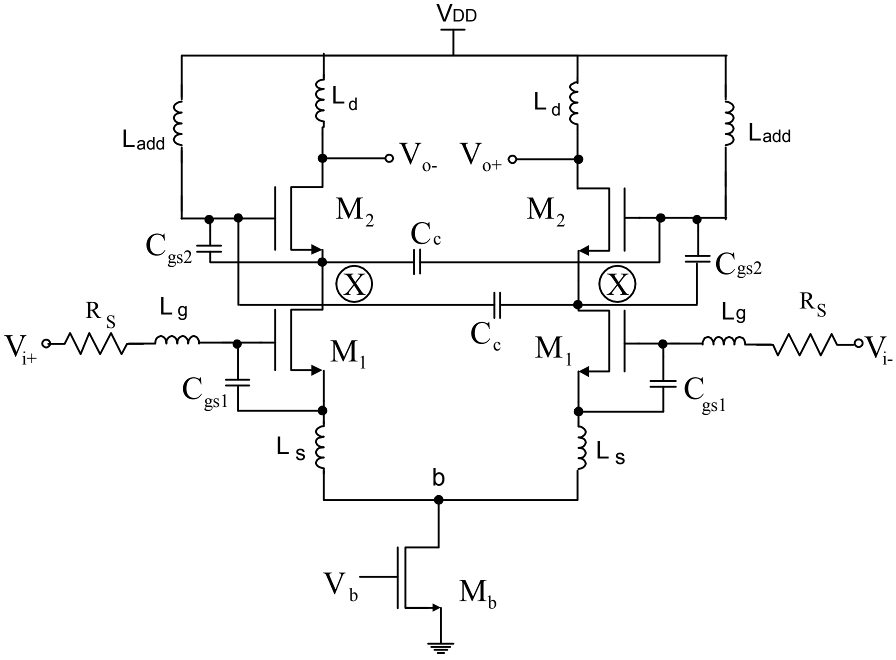


Fig. 3.4. The inductor combined with capacitive cross-coupling technique in a fully-differential cascode CS-LNA.

where F_1 is the LNA noise factor when ignoring the noise contribution from the cascode stage, and $F_c^{'}$ is the noise from the cascode. Note that $B_{eff}^{'}$ is a function of ω.

Since the effect of the parasitic capacitance at node X is cancelled as shown in (3.11)-(3.13), the noise of the cascode transistors is negligible.

The inductor L_{add} can be implemented with either on-chip inductor or bonding wire inductor. Its value is reduced by a factor of 4 with respect to the typical inductor based technique [58]-[60]. Here L_{add} is implemented as a bonding wire inductor. Since now the gates of M_2 are connected out of the chip using the bonding wire inductor, it is desired to add ESD protection structures for these pads. In this design, to verify the proposed concept and to get the optimal results, there are no ESD protection structures for these pads. If the ESD protection circuit is used, it can be modeled in first order as a grounded capacitor parallel with the bonding wire inductor. The parallel LC network should be used to replace the L_{add} in the analysis used in this chapter.

The proposed LNA topology shown in Fig. 3.4 is designed with TSMC 0.35µm CMOS technology and the noise performance is shown in Fig. 3.5. L_g is an ideal inductor, while L_s and L_d are on-chip spiral inductors, which are modeled as pi model using ASITIC software [62]-[64].

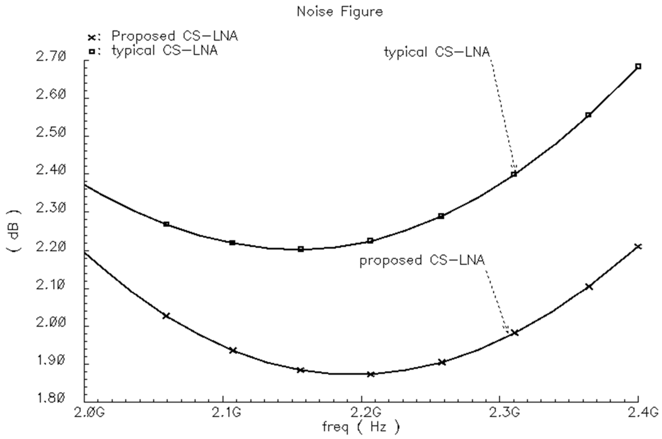


Fig. 3.5. Simulation results of the differential cascode CS-LNA with and without L_{add} and C_c.

The proposed technique reduces the differential cascode CS-LNA NF by 15.8%, i.e. from 2.22dB to 1.87dB at 2.2GHz, which is our designed resonant frequency. However, thanks to the finite Q of the LC tank, noise reduction can be observed over a frequency range. It will be more significant for the LNAs working at higher frequency.

At the lower frequency, $\mathrm{L_{add}}$ short circuited the gates of the cascode transistors to $\mathrm{V_{DD}}$ supply (AC ground). In that case, the total capacitive effects at node X in Fig. 3.4 are not zero and the LNA has worse noise performance.

The bonding wire inductance has different PVT values. From (3.10)-(3.14), at the operating frequency, we obtain that the variations of $G_{m,eff}^{'}$, $B_{eff}^{'}$ and F' can be approximated as:

$$\Delta \mathrm{G}_{\mathrm{m,eff}}^{'} = \mathrm{G}_{\mathrm{m,eff}}^{'}(\mathrm{L_{add}} + \Delta \mathrm{L_{add}}) - \mathrm{G}_{\mathrm{m,eff}}^{'}(\mathrm{L_{add}}) \cong 0 \tag{3.15}$$

$$\Delta \mathrm{B}_{\mathrm{eff}}^{'} = \mathrm{B}_{\mathrm{eff}}^{'}(\mathrm{L_{add}} + \Delta \mathrm{L_{add}}) - \mathrm{B}_{\mathrm{eff}}^{'}(\mathrm{L_{add}}) \cong \frac{1}{\omega_{\mathrm{o}} \mathrm{L_{add}}} \times \frac{\Delta \mathrm{L_{add}}}{\mathrm{L_{add}}} \tag{3.16}$$

$$\begin{aligned} \Delta F^{'} &= F^{'}(L_{add} + \Delta L_{add}) - F^{'}(L_{add}) \cong 4R_s \gamma_2 g_{do2} \left(\frac{\omega_o \Delta B_{eff}^{'}}{\omega_T G_{m,eff}^{'}} \right)^2 \\ &= 4R_s \gamma_2 g_{do2} \left(\frac{\omega_o (4C_{gs2} + C_{sb2} + C_{gd1} + C_{db1})}{\omega_T G_{m,eff}^{'}} \right)^2 \times \left(\frac{\Delta L_{add}}{L_{add}} \right)^2 \end{aligned} \tag{3.17}$$

From (3.15)-(3.17), as an example, with 10% variation in L_{add} value, the proposed technique can still achieve around 96% noise reduction for the cascode device, assuming the ideal L_{add} can entirely eliminate the cascode transistor noise contribution. The noise performance of the LNA with varied inductor L_{add} value (from 3nH to 5nH) is shown in Fig. 3.6. The NF varied from 1.87dB to 1.95dB, that is 4.2% variation for a 67% variation of Ladd.

The LNA NF varies with temperature. The noise reduction with the proposed technique through temperature variation is summarized in Table 3.1. Since the noise of the transistor increases with the increasing temperature, the absolute value of the cascode transistor noise contribution also increases. Thus if it is ideally eliminated, the absolute noise reduction value becomes larger at higher temperature.

Table 3.1. NF improvement versus Temperature

	-45°C	27 °C	85 °C
NF without proposed technique	1.59dB	2.22dB	3.22dB
NF with proposed technique	1.42dB	1.87dB	2.4dB
NF improvement	0.17dB(11%)	0.35dB(16%)	0.82dB(25%)

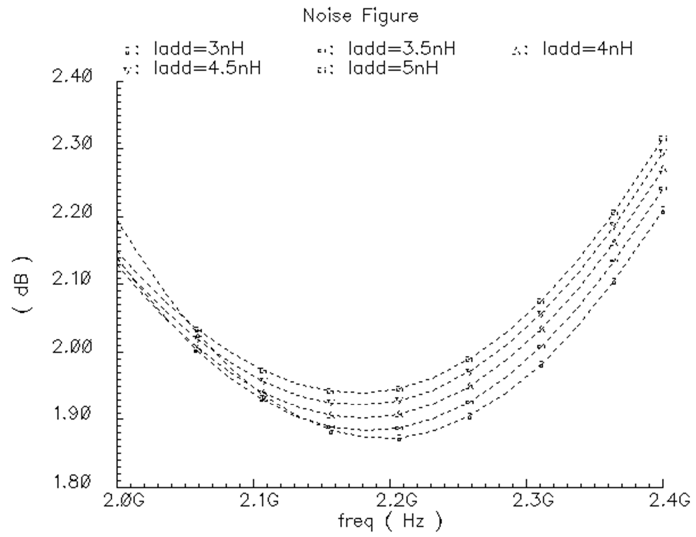


Fig. 3.6. Simulated NF of the differential cascode CS-LNA with the inductor L_{add} value varied from 3nH to 5nH.

3.4 LNA Linearity Improvement with the Proposed Technique

The LNA linearity is normally dominated by the voltage to current conversion transistor in CS stage. If the voltage gain of the first stage is greater than one, the second stage linearity plays a more important role [62]. Since the cascode CS-LNA can be treated as a CS-CG two stage amplifier, the linearity of the proposed topology is analyzed in two parts: 1) the linearity of the first voltage to current conversion stage; 2) the linearity of the cascode stage.

The linearity of the common source MOS transistor or common emitter bipolar transistor is well reported in the literature [9][10][14][20][24][25]. The linearity of the first voltage to current conversion stage is analyzed based on Fig. 3.7.

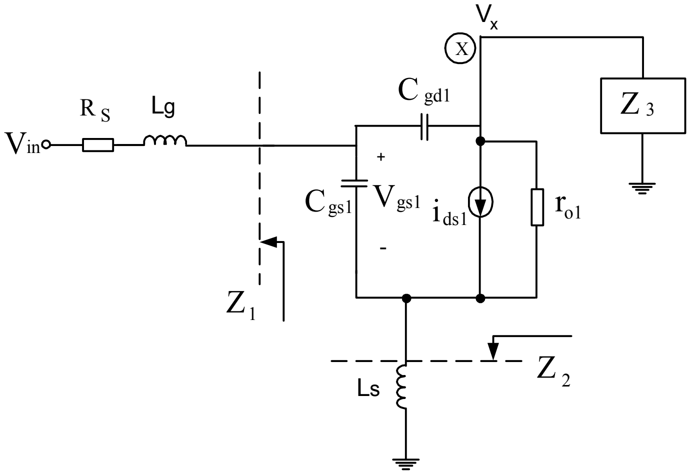


Fig. 3.7. Analyzed CS stage of cascode CS-LNA equivalent circuit.

The drain currents of M1 and M2 in Fig. 3.4 can be expressed as below up to 3rd order:

$$i_{ds} \approx I_{DC} + g_m V_{gs} + g_2 V_{gs}^{\ 2} + g_3 V_{gs}^{\ 3} \tag{3.18}$$

The IIP3 of the first voltage to current conversion stage can be derived using Volterra series as (see Appendix B.4) [9], [10].

$$\mathrm{IIP}_3 = \frac{1}{6R_s \cdot |H(\omega)| \cdot |A_1(\omega)|^3 \cdot |\varepsilon(\Delta\omega, 2\omega)|} \tag{3.19}$$

$$\varepsilon(\Delta\omega, 2\omega) = g_3 - g_{oB} \tag{3.20}$$

$$g_{oB} = \frac{2}{3} g_2^2 [\frac{2}{g_{m1} + g(\Delta\omega)} + \frac{1}{g_{m1} + g(2\omega)}] \tag{3.21}$$

$$g(\omega) = \frac{1 + j\omega C_{gd}[Z_1(\omega) + Z_3(\omega)] + j\omega C_{gs}[Z_1(\omega) + Z_x(\omega)]}{Z_x(\omega)} \tag{3.22}$$

$$Z_x(\omega) = Z_2(\omega) + j\omega C_{gd}[Z_1(\omega)Z_2(\omega) + Z_1(\omega)Z_3(\omega) + Z_2(\omega)Z_3(\omega)] \tag{3.23}$$

where ω is the center frequency of two input tones: ω_1 and ω_2, $\Delta\omega = |\omega_1 - \omega_2|$, $|H(\omega)|$ relates the equivalent input IM3 voltage to the IM3 response of the drain current non-linear terms, $A_1(\omega)$ is the linear transfer function from the input voltage V_{in} to the gate-source voltage V_{gs1}. $Z_1(\omega)$ and $Z_2(\omega)$ are shown in Fig. 3.4. $\varepsilon(\Delta\omega, 2\omega)$ shows the nonlinear contributions from the second and third order terms described in (3.18). For a MOS transistor, it can be found that g_3 and g_{oB} have opposite signs. From (3.19)-(3.20), the reduction of both g_3 and g_{oB} is needed to improve the IIP3.

Z_3 is the impedance looking out of the drain of the main transistor M_1. For the conventional cascode CS-LNA [7], [51]-[52], its relation with the cascode transistor M_2 is described as:

$$Z_3(\Delta\omega) \approx \frac{1}{g_{m2}} \tag{3.24}$$

$$Z_3(2\omega) \approx \frac{1}{g_{m2} + j2\omega C_{gs2}} \tag{3.25}$$

From (3.10)-(3.11), for our proposed LNA, the above values become

$$Z_3^{'}(\Delta\omega) = \frac{1}{G_{m,eff}^{'}(\Delta\omega) + jB_{eff}^{'}(\Delta\omega)} \approx \frac{1}{g_{m2}} \qquad (3.26)$$

$$Z_3^{'}(2\omega) = \frac{1}{G_{m,eff}^{'}(2\omega) + jB_{eff}^{'}(2\omega)} \approx \frac{1}{2g_{m2} + j8\omega C_{gs2}} \qquad (3.27)$$

Z_3 is the same at $\Delta\omega$, and is smaller at 2ω for the proposed LNA. From (3.19)-(3.27), we can find that the proposed LNA reduces the load impedance (Z_3) of the main transistor M_1 and therefore reduces g_{oB} and $\varepsilon(\Delta\omega,2\omega)$, resulting in a higher IIP3.

The linearity of the cascode stage is next analyzed based on Fig. 3.8, where currents $i_{ds2} = g_{m2}(V_2 - V_x)$, i_{1+} and i_{1-} are the differential input signals and i_{d+} and i_{d-} are the differential output signals.

For the cascode stage without the proposed technique, we can express i_d as

$$i_d = i_1 - g(\omega) \cdot V_{gs2} \qquad (3.28)$$

where i_1 is the differential input current ($i_{1+} - i_{1-}$), i_d is the differential output current ($i_{d+} - i_{d-}$), V_{gs2} is the gate-source voltage of the cascode transistor, and

$$g(\omega) = j\omega C_{gs2} \qquad (3.29)$$

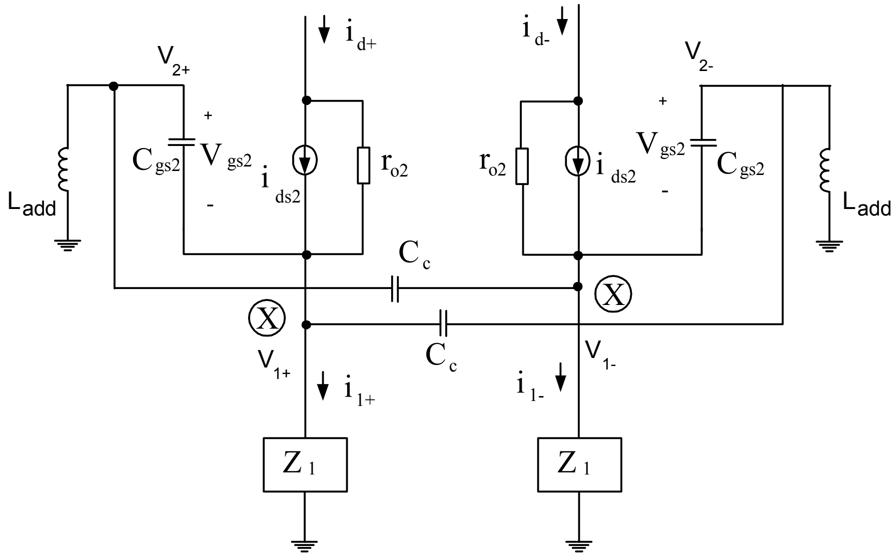


Fig. 3.8. Analyzed cascode stage equivalent circuit.

From (3.28)-(3.29), due to C_{gs2}, the nonlinearity of transistor M_2 influences the overall linearity of the LNA. The A_{IIP3} of the conventional cascode stage without L_{add} and C_c can be derived using Volterra series as

$$A^2_{IIP_3} = \frac{4}{3} \cdot \frac{1}{|H(\omega)| \cdot |A_1(\omega)|^3 \cdot |\varepsilon(\Delta\omega,2\omega)|} \qquad (3.30)$$

$\varepsilon(\Delta\omega,2\omega)$ and g_{oB} are defined the same as in (3.20) and (3.21).

$$H(\omega) = \frac{g(\omega)}{g_{m1}} \qquad (3.31)$$

$$A_1(\omega) = \frac{1}{g_{m1} + g(\omega)} \qquad (3.32)$$

For the cascode stage with the proposed technique, we can obtain:

$$i_d = i_1 - g'(\omega)\cdot V_{gs2} \tag{3.33}$$

$$g'(\omega) = \frac{4j\omega C_{gs2}\cdot j\omega C_c + \frac{1}{j\omega L_{add}}(j\omega C_{gs2} + j\omega C_c)}{2j\omega C_c + \frac{1}{j\omega L_{add}}} + \omega C_{sb2} + \omega C_{gd1} + \omega C_{db1} \tag{3.34}$$

Note that in (3.33), the current flowing into C_c is included. If $\omega C_c >> \omega C_{gs2}$, $\omega C_c >> \omega C_{gs2} - \frac{1}{\omega L_{add}}$ and inductor L_{add} resonates with the effective capacitance at node X at $\omega = \omega_o$, (3.34) becomes

$$g'(\omega) \approx \frac{1}{j\omega_o L_{add}} + 4j\omega_o C_{gs2} + \omega_o C_{sb2} + \omega_o C_{gd1} + \omega_o C_{db1} \approx 0 \tag{3.35}$$

and (3.33) yields

$$i_d = i_1 - g'(\omega_o)\cdot V_{gs2} \approx i_1 \tag{3.36}$$

Thus according to (3.36), there is no linearity degradation from the cascode stage. The A_{IIP_3} of the cascode stage with the proposed technique has the same expression as (3.30) but with different g(ω) as defined by (3.34). From the simulation, the proposed technique increases the linearity by 2.35dBm as shown in Fig. 3.9.

From (3.33)-(3.36), the inductor L_{add} can resonate with the effective capacitance at node X to completely remove the nonlinearity contribution from the cascode transistor M_2. The linearity improvement will vary with different L_{add} values due to the PVT variation. The IIP3 of LNA is shown in Table 3.2. It varied less than 1.2dBm with inductor value varied from 0% to 10%.

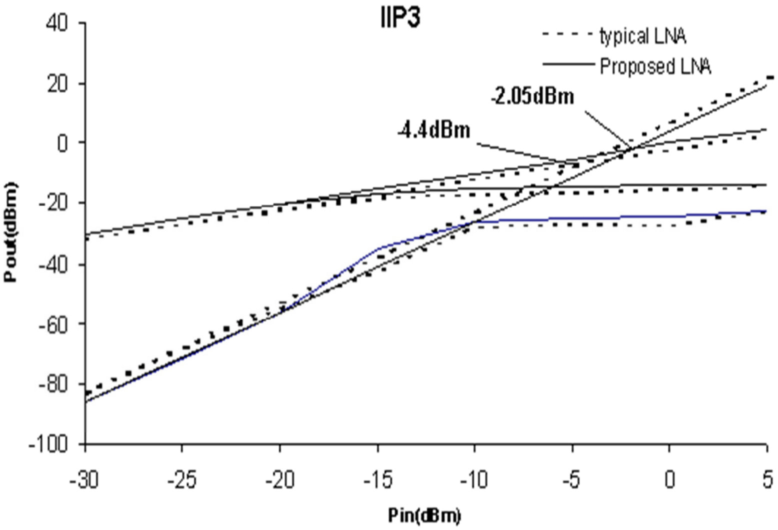


Fig. 3.9. IIP3 of the differential cascode CS-LNA with and without L_{add} and C_c.

Table 3.2. IIP3 versus L_{add}

	Typical LNA	Proposed LNA with varied L_{add}		
		3nH (0%)	3.15nH (5%)	3.3nH (10%)
IIP3(dBm)	-4.4	-2.05	-2.3	-2.5

For the proposed cascode LNA topology shown in Fig. 3.4, we can draw the conclusion that the capacitive cross-coupling technique improves the linearity by increasing the effective transconductance of the cascode stage (M2), thus reducing the load impedance of the main transistor M_1. Therefore, the reduced voltage swing at node X (drain of M1) improves the linearity of CS stage of the LNA. The inductor L_{add}

resonates with the parasitic capacitance at node X and therefore eliminates the nonlinearity and noise contribution from the cascode stage.

3.5 Effects of the Technique on the LNA S11, Voltage Gain and LNA Stability

3.5.1 Effect on the LNA S11

For the typical cascode CS-LNA, C_{gd1} of the transistor M_1 reflects Miller impedance at the gate of M_1. However it is not purely capacitive and its susceptance yields

$$B_{mil1}(j\omega) = (1 + Av(j\omega)) \times sC_{gd1} = \frac{j\omega C_{gd1} g_{m1}}{1 + j\omega g_{m1} \frac{L_s}{C_{gs1}} - L_s C_{gs1}} \times \frac{1}{g_{m2} + j\omega C_x} \qquad (3.37)$$

where $Av(j\omega)$ is the voltage gain from the gate to the drain of M1, and C_x is defined in (3.2). For the proposed LNA, it changes to

$$B_{mil1}{}'(j\omega) = (1 + Av'(j\omega)) \times sC_{gd1} = \frac{j\omega C_{gd1} g_{m1}}{1 + j\omega g_{m1} \frac{L_s}{C_{gs1}} - L_s C_{gs1}} \times \frac{1}{G'_{m,eff} + jB'_{eff}} \qquad (3.38)$$

where $G'_{m,eff}$ and B'_{eff} are defined in (3.10)-(3.12). According to (3.37)-(3.38), since the effective transconductance of the cascode stage increases, the gain of the first stage reduces, which leads to a reduced Miller effect of C_{gd1} of transistor M_1. Therefore the input matching is not very sensitive to the variations of the inductor L_{add}. According to Fig. 3.10, the input resonant frequency varied less than 1% for the L_{add} value varied 66%, which is from 3nH to 5nH.

3.5.2 Effect on the LNA Voltage Gain

Under the input impedance matched condition, the voltage gain of the inductively degenerated cascode CS-LNA can be derived from Fig. 3.1 and Fig. 3.2

$$A_v(j\omega) = g_{m1}\frac{1}{2R_s\omega_o C_{gs1}}\frac{g_{m2}}{\sqrt{(g_{m2})^2+(\omega_o C_x)^2}}Z_o = g_{m1}Q_{in}Z_o\frac{g_{m2}}{\sqrt{(g_{m2})^2+(\omega_o C_x)^2}}$$

(3.39)

where $Q_{in} = \frac{1}{2R_s\omega_o C_{gs1}}$ is the quality factor of the LNA input network and Z_o is the overall output impedance. With the proposed technique, the cascode CS-LNA gain of Fig. 3.4 becomes

$$A_v(j\omega) = g_{m1}Q_{in}Z_o\frac{G^{'}_{m,eff}}{\sqrt{(G^{'}_{m,eff})^2+(B^{'}_{eff})^2}} \tag{3.40}$$

$G^{'}_{m,eff}$ and $B^{'}_{eff}$ are defined in (3.10)-(3.12).

The gain of the designed fully-differential CS-LNA is shown in Fig. 3.11, where the LNA drives 50Ω resistor. According to (3.39)-(3.40) and simulation results in Fig. 3.11, the proposed technique increases the overall LNA gain by around 2dB(25%).

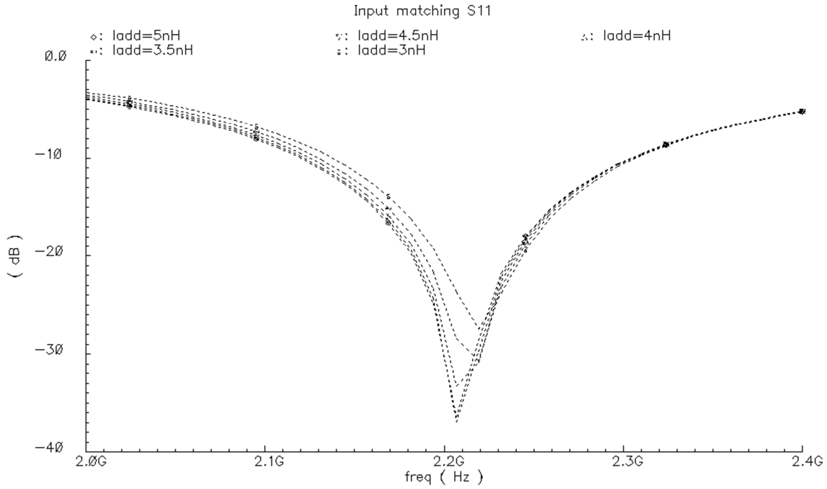


Fig. 3.10. Simulated S11 of the differential cascode CS-LNA with the inductor L_{add} value varied from 3nH to 5nH.

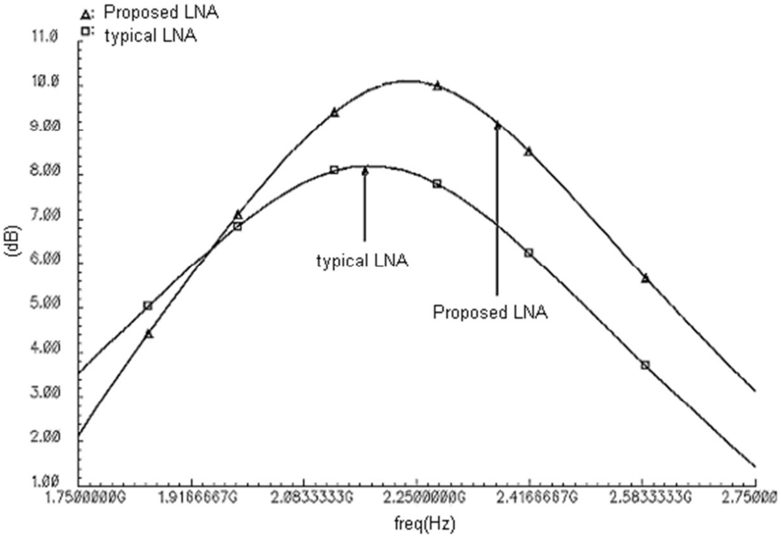


Fig. 3.11. Voltage gain simulation results of the fully-differential cascodeCS-LNA with and without L_{add} and C_c.

In most of the wireless transceivers, the following stage of the LNA is a mixer. It is a capacitive load rather than a 50Ω load, which is the case in this simulation. The source follower can drive the off-chip 50Ω with the voltage gain around 1. A source follower buffer is added after the LNA to drive a 50Ω load. This testing setup of the LNA voltage gain is shown in Fig. 3.12 where LNA drives a buffer. Fig. 3.13 is the simulated LNA voltage gain. The LNA is simulated with a source follower to drive the off-chip 50ohm and the voltage gain of interest is investigated before the source follower. We used an ideal balun in the simulation. In practice, the LNA directly drives a practical balun without the source follower buffer. The noise and gain influence of the balun is de-embeded. In this way, we can estimate the LNA voltage gain while driving the mixer in the wireless receiver. Since the buffer provides a 250fF capacitive load (C_{LOAD}) rather than 50Ω resistive load, the LNA voltage gain increases to 20.4dB as shown in Fig. 3.13.

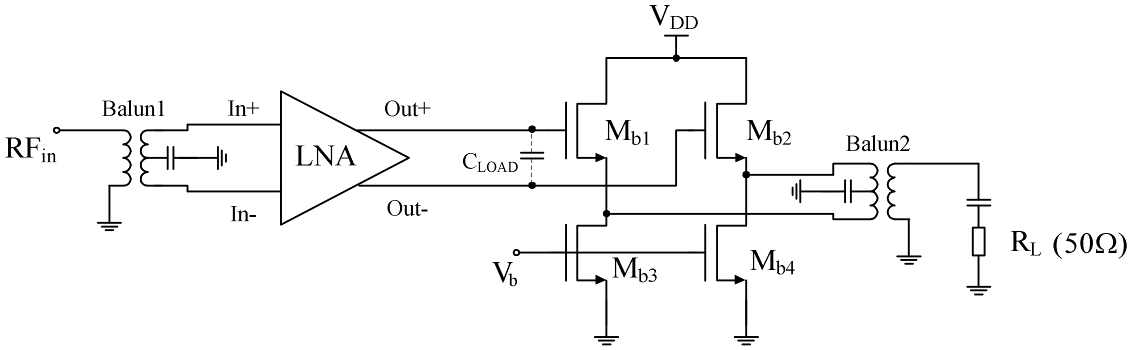


Fig. 3.12. Voltage gain testing setup of the fully-differential cascode CS-LNA when driving the on-chip buffer.

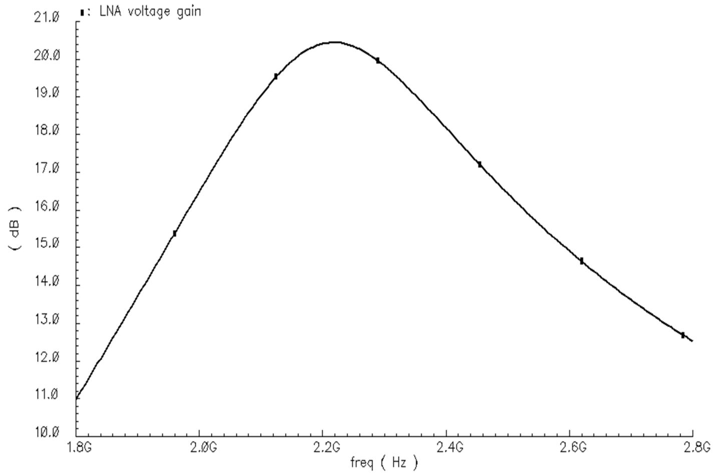


Fig. 3.13. Voltage gain simulation results of the fully-differential cascode CS-LNA when driving an on-chip buffer.

3.5.3 Effect on the LNA Stability

In addition to gain, NF, and linearity, the stability of LNAs is also an important design parameter. When feedback paths exit from the output to the input, the LNA may become unstable in these three situations: 1) with certain combinations of source and load impedances; 2) with process, voltage, and temperature variations; 3) operating at the extreme frequencies. Therefore, a stability factor [7] is defined to characterize LNAs:

$$K = \frac{1 + |\Delta|^2 - |S11|^2 - |S22|^2}{2|S21||S12|} \tag{3.41}$$

where $\Delta = S11S22 - S12S21$. The unconditional stability requirement, i.e. the LNA does not resonate with any combinations of source/load impedances, is K>1 and $|\Delta| < 1$ at a wide frequency range. When the input and output of the LNA are matched to the source and load impedance, S11 and S22 are almost 0. With the decreasing of the S12, $|\Delta|$ reduces, which means the better stability of the LNA. The S12 reflects the input output isolation of the LNA. Compared with the typical LNA, the added inductor L_{add} at the gate of the cascode transistor M2 along with the inherent capacitances provides a low impedance path for the output signal feedback to the input, which helps to improve the input output isolation(S12)[65]. The cross-coupling capacitor Cc forms a signal path from the gate of the cascode transistor M2 to the source of M2, which reduces the isolation effect of the transistor $M_{2.}$ The proposed technique presents an overall comparable isolation effect with the typical LNA with around 3dB worse S12 value in the simulation. From simulation, the K value of the LNA is 52.5 at 2.2GHz without Ladd and Cc. K becomes 30.5 at 2.2GHz with Ladd and Cc. The difference of the K value is partly due to the 2dB S21 difference and 3dB S12 difference with/without Ladd and Cc. The LNA is stable in both cases.

3.6 Design and Measurement Results

A fully-differential cascode CS-LNA was designed and fabricated using a proposed inclusive noise reduction and linearity improvement technique. The inductor L_g is an off-chip inductor. The added inductor L_{add} (around 3nH) is a bonding wire inductor. The inductors L_s (0.5nH) and L_d (3nH) are on-chip spiral inductors, with $Q \approx 3$. The design was implemented using TSMC 0.35 μm CMOS technology. The chip micrograph is shown in Fig. 3.14. The LNA occupies 1300μm×1000μm active area, with the LNA core using 850μm×850μm active area.

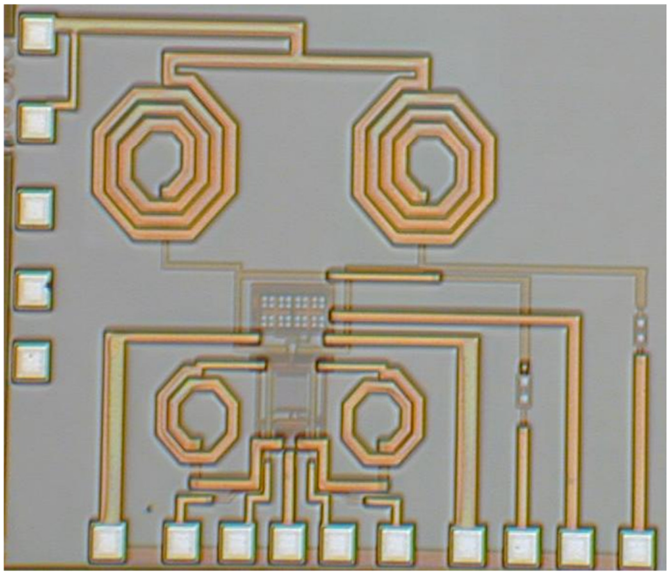


Fig. 3.14. Chip micrograph of the differential cascode CS-LNA.

Fig. 3.15, Fig. 3.16, and Fig. 3.17 show the lab measurement setup for LNA S parameter, NF, and linearity, respectively. Before measuring the S parameters using the network analyzer, we should first perform a "full two port" calibration within the desired frequency range to take away the cabling effects. For the NF testing, we first used a noise source to calibrate the loop, then insert the LNA test board into the loop. Note that if the LNA gain is too low, an off-the-shelf commercial LNA can also be inserted to boost the gain thus reduce the noise effect from equipments and improve the measurement accuracy. The noise from the commercial LNA can be de-embedded using the cascaded NF equation. As for the high linearity testing, we should first characterize the attenuation and distortion from cable and power combiner before taking the IIP2 and IIP3 data of the LNA.

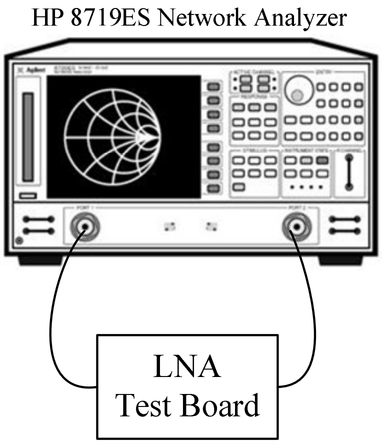
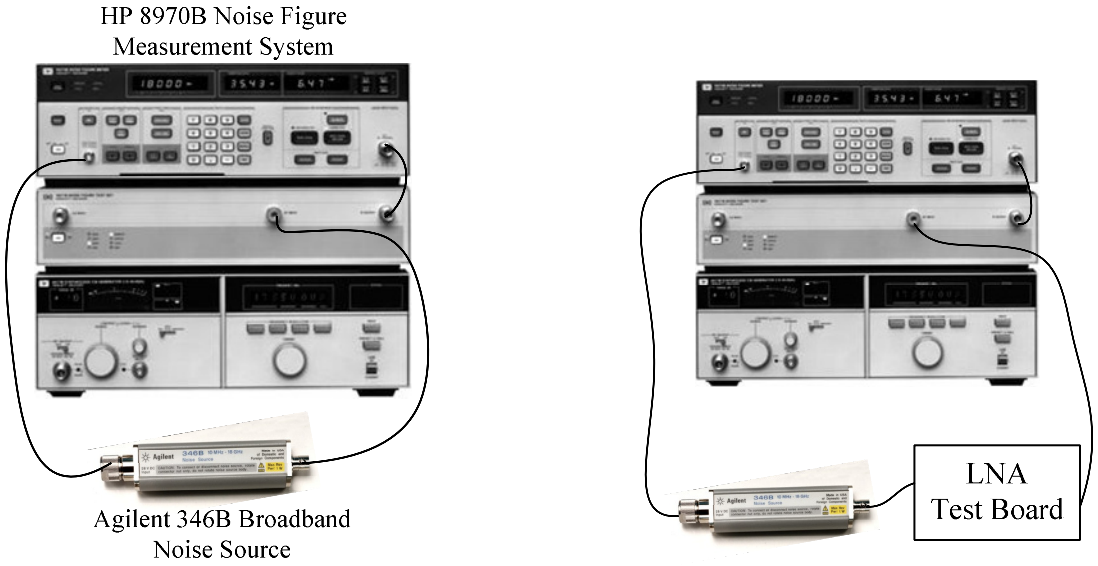


Fig. 3.15. LNA S parameter measurement.



(a)

(b)

Fig. 3.16. LNA noise figure measurement (a) instrument calibration (b) measurement.

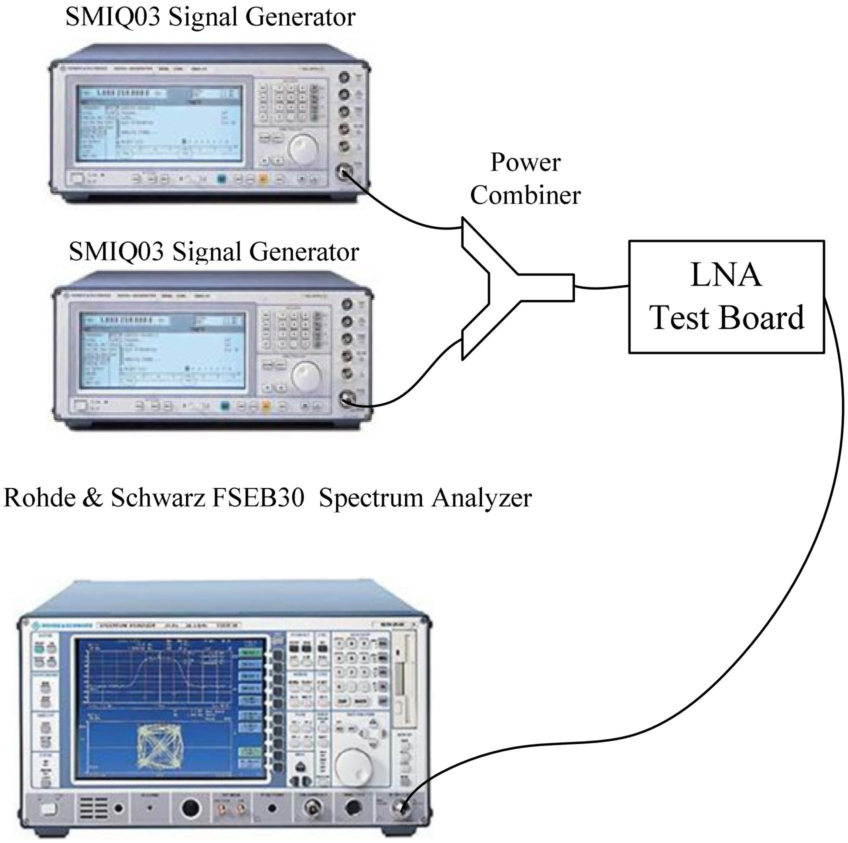


Fig. 3.17. LNA linearity (IIP2, IIP3) measurement.

L_g value is adjusted in the measurement to achieve the input impedance matching at the desired frequency. Fig. 3.18 shows the measured S11, S21 and S12. The LNA power gain is 8.4dB at 2.2GHz. If followed by a buffer, the LNA output impedance is larger than 50Ω and the LNA gain increases up to 20.4dB in simulation. S11 is less than -13 dB. And S12 is less than -30dB. Fig. 3.19 shows the measured NF of the LNA. The LNA has 1.92dB NF. The third-order input intercept point (IIP3) was measured using a two-tone test: 2.2GHz and 2.22GHz. It is shown in Fig. 3.20. The IIP3 is -2.55dBm. The core LNA draws 9mA from a 1.8V power supply. Due to the mismatch of the gate inductor Ladd, the noise of the power supply can inject into the LNA output. The PSRR of the LNA with 5% and 10% Ladd mismatches is shown in Fig. 3.21. The LNA has better than -24dB PSRR at 2.2GHz with 10% Ladd mismatch.

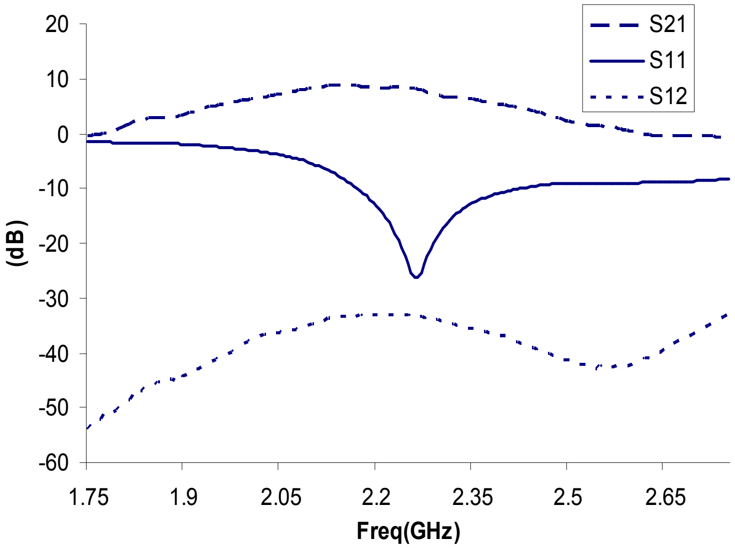


Fig. 3.18. Measured S11, S12 and S21 of the differential cascode CS-LNA.

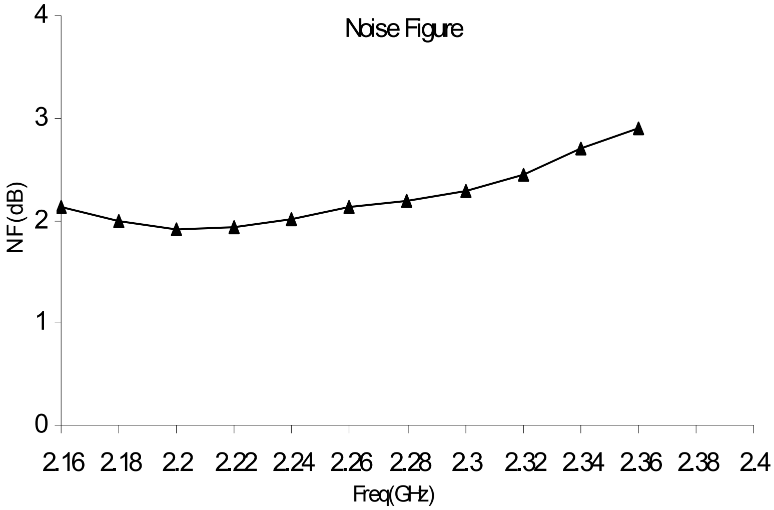


Fig. 3.19. Measured NF of the differential cascode CS-LNA.

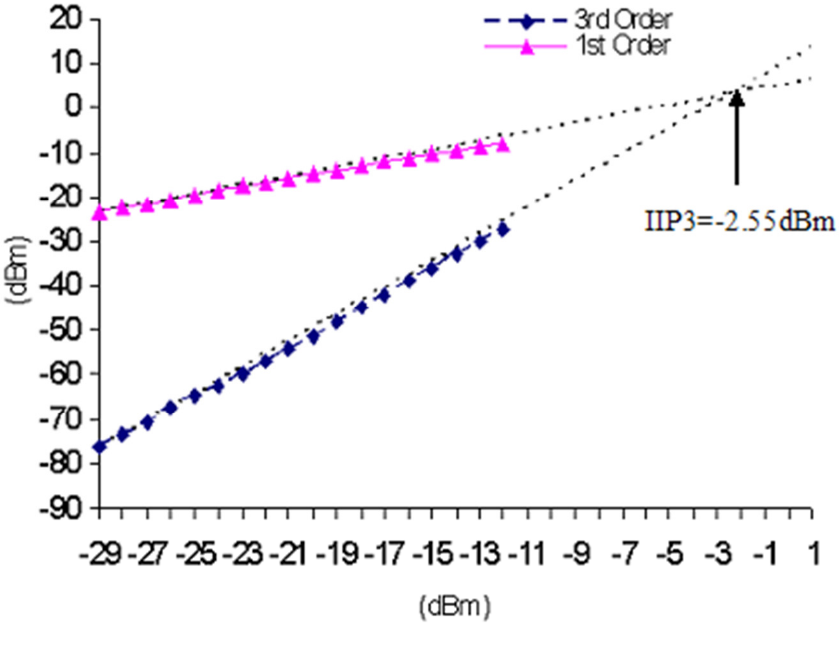


Fig. 3.20. Measured IIP3 of the differential cascode CS-LNA, with two tones at 2.2GHz and 2.22GHz.

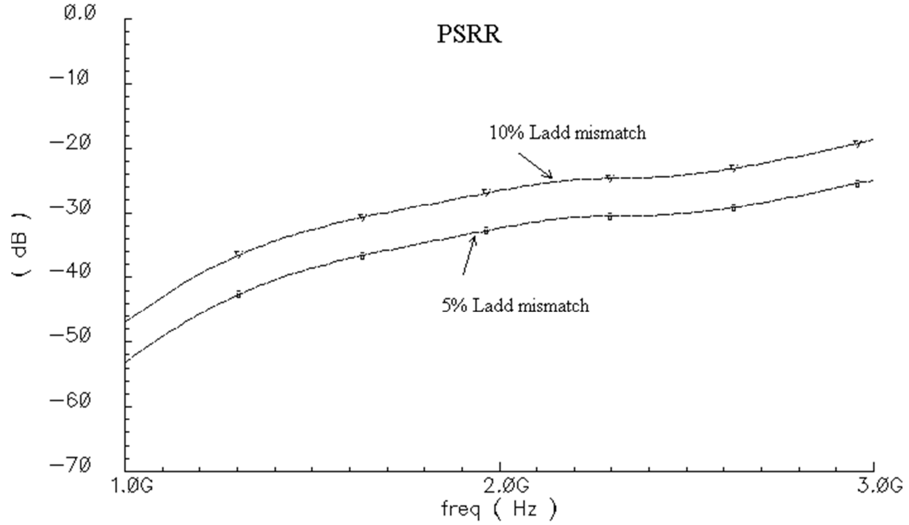


Fig. 3.21. PSRR of the LNA with 5% and 10% Ladd mismatches.

Table 3.3. Performances compared with the prior published cascode CS-LNAs

Parameters	[66]	[67]	[68]	[69]	This work	
					Simulated	Measured
Frequency(GHz)	2.45	2.46	2.40	0.95	2.2	2.2
S11(dB)	<-14.2	<-18.4	<-33	<-14	<-13	<-13
S21(dB)	15.1*	14	6	17	10 (20.4)+	8.6
NF(dB)	2.88	2.36	4.80	3.40	1.87	1.92
IIP3 (dBm)	2.20	-2.20	0.55	-5.10	-2.05	-2.55
Bias current(mA)	8.1	3.1	N/A	5.6	4.5×2	4.5×2
Power supply(V)	3	1.5	3.3	2.3	1.8	1.8
Topology	Single-ended	Single-ended	Single-ended	Single-ended	Fully-differential	Fully-differential
CMOS Process	0.25µm	0.15µm	0.35µm	0.35µm	0.35µm	0.35µm

*: In fact in [66] they reported the transducer gain.

+: 20.4 dB is obtained when an output buffer is used instead of 50Ω load.

The comparison of this LNA with the published literatures is summarized in Table 3.3. Although the designed LNA is a fully-differential structure in 0.35μm process, it provides the best noise performance. The published LNAs consume less bias current because of the single-ended structure and more advanced technology. The linearity in [66] is higher due to the larger bias current and more voltage headroom for the transistors. Although the current source of the designed fully-differential LNA reduces the voltage headroom, it still achieves comparable linearity with respect to [67]. The LNA gain is proportional to the inductor quality factor and the inductor value as shown below [67]

$$\mathrm{Gain} \propto \mathrm{R_p} \propto \mathrm{Q_d^2 R_d} \propto \omega_o \mathrm{Q_d L_d} \tag{3.42}$$

where R_d is the series resistance of L_d, R_p is the parallel resistance of L_d obtained from the series to parallel transformation, and Q_d is the quality factor of L_d. The LNA is designed in 0.35μm process with a low Q on-chip inductor, which results in a smaller gain. After adding a buffer (with similar input impedance of a typical CMOS Gilbert Cell) after the LNA, the LNA can achieve around 20.4dB voltage gain, which is sufficient for a number of wireless applications.

3.7 The Effectiveness of the Proposed Technique in Deep Sub-micron Process

In the deep sub-micron process, the parasitic capacitance of the devices is smaller, thus its effect explained in this paper becomes less significant at the lower operating frequency, but as the operating frequency increases to such as 10GHz or higher, the same effect will appear even in the advanced process. On the other hand, the output impedance of the transistor is smaller in the advanced process, which increases the noise contribution of the cascode transistor. This effect combined with the parasitic capacitance makes the cascode transistor to be still an important noise contributor. The proposed technique can still be effective under these conditions, and the theoretical analysis is also valid. The proposed solution applies the capacitive cross-coupling technique to the cascode transistor of the LNA, which can increase the effective transconductance of the cascode transistor and improve the linearity of the common source stage of the LNA. The gate inductor effectively combined with the cross-coupling capacitor can reduce the noise and the nonlinearity influence of the cascode transistor with a smaller inductor value as proved in section 3.3 and 3.4.

For the LNA working at low frequency in deep sub-micron process, the proposed technique requires a large gate inductor Ladd value due to the smaller parasitic capacitance, thus is not suitable for integration. In that case, the proposed technique has its own limitation. However, we can observe from equation (3.7) that in advanced process, ω_T increases and C_x decreases, so when the operating frequency ω_o increases to such as 10GHz or higher, the noise contribution from cascode device(Fc) becomes comparable to F1 and the noise reduction inductor Ladd value also reduces, enabling easy integration, thus our proposed technique can be applied to the LNA to reduce Fc. To verify our proposed technique in the deep sub-micron process, the LNA is re-designed in UMC 0.13μm CMOS process and simulated based on the noise model provided by UMC. At 10GHz, the proposed technique reduces the differential cascode CS-LNA NF from 1.55dB to 0.95dB, with L_{add} value as 0.5nH, as shown in Fig. 3.22.

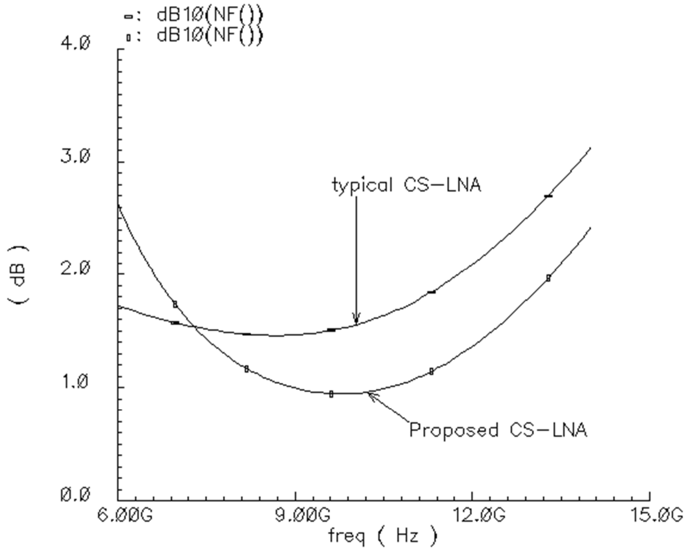


Fig. 3.22. NF simulation results of the differential cascode CS-LNA with and without L_{add} and C_c in UMC 0.13μm CMOS process.

3.8 Conclusions

In this chapter, a linearity improvement and noise reduction technique for a differential cascode CS-LNA was proposed. The inductor connected at the gate of the cascode transistor and the capacitive cross-coupling are strategically combined to reduce the noise and nonlinearity contributions of the cascode transistors. It is the first time that the capacitive cross-coupling technique is applied to the cascode transistors of the CS-LNA. It increases the effective transconductance of the cascode transistor, reduces the impedance seen out by the drain of the main transistor, and thus improving the linearity of the CS stage in the LNA. The inductor L_{add} resonates with the effective capacitance at the drain node of the main transistor with smaller inductance value compared with the typical inductor based technique. It ideally removes the noise and linearity influences from the cascode transistor, and results in a higher voltage gain. The proposed technique is theoretically formulated. From simulation results in TSMC 0.35µm CMOS process, it reduces the LNA NF by 0.35dB at 2.2GHz, and improves the LNA IIP3 by 2.35dBm. To illustrate the use of the proposed approach in small size technology, a10GHz LNA is also designed using UMC 0.13µm CMOS process. The proposed technique reduces the NF from 1.55dB to 0.95dB, which is simulated based on the noise model provided by UMC. This verifies the validity of our proposed technique in the deep sub-micron process.

CHAPTER IV

PROPOSED LINEARIZATION TECHNIQUE FOR AN ULTRA-WIDEBAND LNA*

4.1 Introduction

Growing research on reconfigurable multi-band/multi-standard and ultra-wideband (UWB) transceivers has sparked increased interest in broadband LNA design. A broadband LNA must provide good input matching, high linearity, and low noise figure (NF) over a multi-GHz bandwidth (BW), while consuming little power and die area. To implement broadband impedance matching, a bandpass-filter-(BPF-) based, inductively degenerated common-source (CS) CMOS LNA and a SiGe common-emitter LNA have been proposed in [70] and [71], respectively. The BPF-based UWB CG-LNA first proposed in [72] reduces power and improves the linearity compared to the UWB CS-LNA. However, the large number of inductors requires large area and increases the NF [70]-[72]. Using a CG transistor for input matching is reported in [29]-[30] , [73]-[74], but the additional CS stage consumes more power and degrades the linearity. A differential UWB CG-LNA employs capacitive cross-coupling to reduce the NF [75], but this cross-coupling also increases the quality factor of the parallel RLC input network, reducing the matching BW.

A big design challenge for UWB LNAs is the stringent linearity requirement

*©[2009] IEEE. Reprinted, with permission, from “A Low-Power, Linearized, Ultra-Wideband LNA Design Technique”, by Heng Zhang, Xiaohua Fan, and Edgar Sánchez-Sinencio, *IEEE Journal of Solid-State Circuits,* vol. 44, no. 2. pp. 320-330, Feb. 2009.

system, and the cross-modulation/inter-modulation caused by blockers or transmitter leakage [29] in a reconfigurable receiver. Furthermore, while f_T increases with technology scaling, linearity worsens due to lower supply voltage and high-field mobility effects [29]. Therefore, wideband linearization in deep-submicron CMOS process is a new trend. However, most of the linearization methods reported so far target applications that are either narrowband or have operating frequencies below 3GHz [9]-[29]. To the authors' knowledge, [30] is the first work to explore linearization technique for wideband LNAs with frequencies up to 6GHz.

A linearization method for high-frequency wideband applications is desired. Optimizing the overdrive voltage (V_{gs}-V_{th}) [16], [75] leads to a linearity boost region for fairly narrow range of input amplitude, and an increased sensitivity to process variation. The feed-forward distortion cancellation technique [17]-[30] extends the linearity improvement region. In [17], a coaxial assembly is required for accurate power splitting which is not feasible for practical applications. The derivative superposition (DS) method [19]-[21], [24]-[25] uses an additional transistor's nonlinearity to cancel that of the main device; it involves MOS transistors working in triode [19] or weak inversion region [14][24][25]; thus, these are mainly effective at relatively low frequencies. A bipolar in CMOS process is used [20] to push the operating frequency to 3GHz. However, the common problem existing in all the reported DS methods is its difficulty to match the transistors working in different regions or match a bipolar with a MOS transistor, resulting in a linearity improvement highly sensitive to PVT variations, and sub-optimal nonlinearity cancellation in practice. The post-distortion method [31]-[32]

uses all transistors in saturation region and also avoids the input matching degradation; however, the two cascode paths will introduce linearity and BW degradation at high frequencies [15], thus more inductors will be needed to avoid gain roll off for wideband application [72].

In this chapter, a single-stage, low-power UWB CG-LNA is introduced, which has the simplest input matching network and the lowest power consumption compared to the prior reported single-ended UWB LNAs. Furthermore, a linearization technique is implemented on the single-stage cascode UWB CG-LNA. The added simple linearization circuitry does not affect the wideband input matching and has minimum power/area overhead. Section 4.2 describes the properties of the typical CS-LNA and CG-LNA. Section 4.3 presents both the proposed single-stage, single-transistor UWB CG-LNA and the cascode (two-transistor) version, and analyzes their noise and linearity. Section 4.4 presents the proposed linearization technique. Theory and simulation are compared, and the impact of PVT variations is discussed. Section 4.5 addresses the effects of the proposed linearization technique on the S_{11} and NF of LNA. Measurement results and conclusions are presented in Section 4.6 and Section 4.7, respectively.

4.2 Properties of the CS-LNA and CG-LNA

Fig. 4.1 and Fig. 4.2 show a typical inductively degenerated common-source LNA (CS-LNA) [7] and a common-gate LNA (CG-LNA), respectively. C_{gs1} is the parasitic gate-to-source capacitance. Their input impedance $Z_{in}(s)$ seen by R_s and the quality factor of the input matching network Q_{match} are listed in Table 4.1. For simplicity, all other parasitics and body effects are ignored.

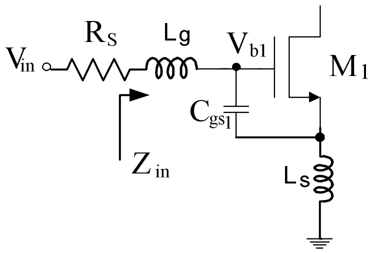


Fig. 4.1. Typical inductor-degenerated common source LNA.

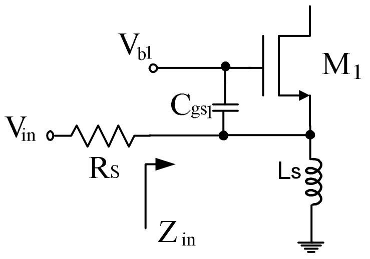


Fig. 4.2. Typical common gate LNA.

A lower Q_{match} results in a wider BW. Due to the relatively high Q of CS-LNAs' matching network, the CS-LNA cannot meet UWB matching requirements without advanced design techniques [70][71].

Table 4.1. CS-LNA versus CG-LNA topologies

Topology	$Z_{in}(s)$ seen from R_s	Q_{match}
CS-LNA	$\dfrac{s^2 + s\dfrac{g_{m1}L_s}{(L_s+L_g)C_{gs1}} + \dfrac{1}{(L_s+L_g)C_{gs1}}}{s/(L_s+L_g)}$	$\dfrac{1}{2\omega C_{gs1}R_s}$
CG-LNA	$\dfrac{s/C_{gs1}}{s^2 + s\dfrac{g_{m1}}{C_{gs1}} + \dfrac{1}{L_sC_{gs1}}}$	$\dfrac{\omega C_{gs1}R_s}{2}$

The CG-LNA, however, has a parallel resonant network with low Q_{match}. For example, C_{gs1} = 0.3pF yields Q_{match}(f=5GHz) = 0.24 and hence BW= 21GHz. Because Q_{match} is proportional to C_{gs1}, Q_{match} will decrease and thus BW will increase as technology scales. Therefore the CG-LNA can easily implement broadband impedance matching without many extra components, dramatically saving area and avoiding on-chip inductor resistive losses [72]-[75]. Besides the simple and robust input matching architecture, the CG-LNA also has better linearity, lower power consumption, and better input-output isolation [72].

The NF of the CS-LNA is generally superior to that of the CG-LNA, because the CG-LNA's NF is limited by 1/gm input matching. However, the CG-LNA provides better noise performance for higher operating frequency ratios ω_0/ω_T, as its noise factor is only a weak function of ω_0/ω_T, while the CS-LNA's noise is proportional to ω_0/ω_T [55]. A typical design in UMC 0.13µm CMOS process shows that: at $\omega_0/\omega_T = 0.2$, the

NF for the CS-LNA and the CG-LNA is 3dB and 5.8dB respectively, but for $\omega_0 / \omega_T > 0.66$, the CG-LNA starts to outperform the CS-LNA in NF, and at $\omega_0 = \omega_T$, the CG-LNA NF is 6.3dB, while the NF of CS-LNA has increased to 7.7dB. Therefore, the CG-LNA has a relatively flat NF over a wide frequency range, thus provides superior performance for broadband applications.

4.3 Proposed Low Power Single-Stage UWB CG-LNA

4.3.1 Design Considerations of the Proposed CG-LNA

This chapter details the design of two single-stage UWB CG-LNAs in 0.13μm CMOS process--one single-transistor and the other two-transistors (cascode). The basic topologies are shown in Fig. 4.3 and Fig. 4.4. C_{gs1} and C_{pad} are the parasitic capacitance of transistor M_1 and the input pad respectively. M_3 and M_4 form a buffer to drive the test equipment and also emulate the input impedance of the mixer. L_s, L_D and L_c are on-chip spiral inductors. L_s, C_{gs1}, C_{pad}, and the equivalent impedance of M_1 form a parallel low-Q resonant network. Proper selection of the resonant frequency and Q matches the input to R_s over the whole BW. Inductor L_D is used to achieve flat gain [7], [70]-[75].

The single-transistor LNA demonstrates the simplest topology for a UWB LNA. Adding transistor M_2 (Fig. 4.4) improves isolation and increases low frequency gain by about 2~3dB; however, the parasitic capacitances of M_2 degrades gain, linearity, and NF at high frequency [15], [77]. Inserting inductor L_c partially compensates this degradation.

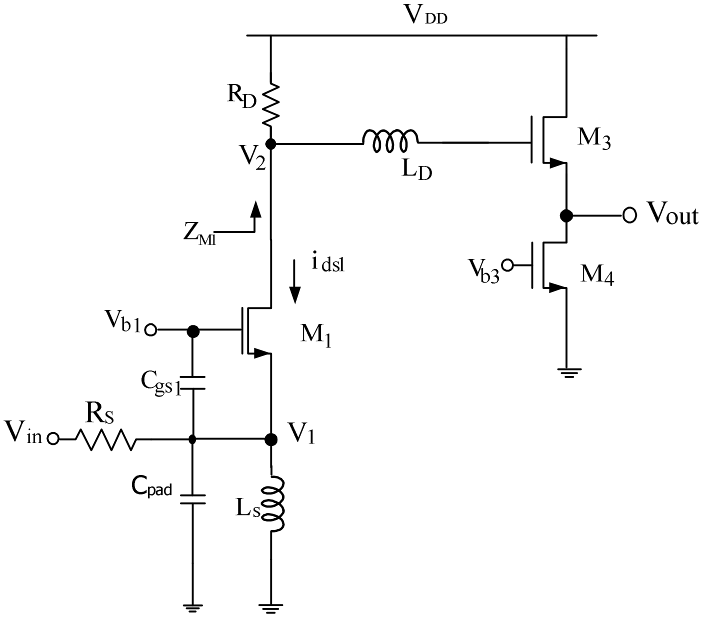


Fig. 4.3. Proposed single-stage single transistor UWB CG-LNA.

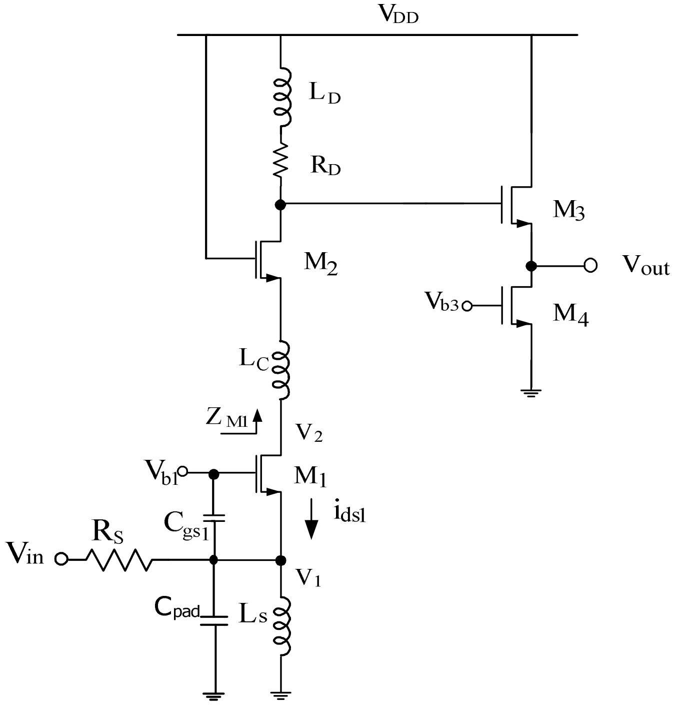


Fig. 4.4. Proposed single-stage cascode UWB CG-LNA.

4.3.2 Noise analysis of the proposed CG-LNA

The overall transconductance of the CG-LNA in Fig. 4.3 and Fig. 4.4 is given by:

$$G_m = \frac{i_{ds1}}{V_{in}} = \frac{\left|Z_{in}(s)\right|}{\left|Z_{in}(s)+R_s\right|} g_{m1} \tag{4.1}$$

where $Z_{in}(s)$ is defined in Table 4.1. The CG-LNA noise factor (neglecting r_o) can be derived as follows.

1) Input referred noise due to M1 channel noise:

$$\frac{4kT\gamma g_{d0}\cdot Z_{load}^2}{Gain_{LNA}} = \frac{4kT\frac{\gamma}{\alpha}g_{m1}\cdot Z_{load}^2}{4kTR_s\cdot g_{m1}^2\cdot Z_{load}^2} = \frac{\gamma}{\alpha}\cdot\frac{1}{g_{m1}R_s} \tag{4.2}$$

where γ, α, and δ are process-dependent parameters[7], and g_{d0} is the drain-source conductance at zero V_{DS}.

2) Input referred noise due to M1 gate noise:

$$\begin{aligned}\frac{4kT\delta g_g\cdot Z_{load}^2}{Gain_{LNA}} &= \frac{4kT\delta\omega^2C_{gs}^2}{5g_{d0}}\cdot\frac{Z_{load}^2}{Gain_{LNA}}\\ &= \frac{4kT\delta\alpha\omega^2C_{gs}^2}{5g_{m1}}\cdot\frac{Z_{load}^2}{4kTR_s\cdot g_{m1}^2\cdot Z_{load}^2} = \frac{\delta\alpha}{5}\cdot\left(\frac{\omega}{\omega_T}\right)^2\cdot\frac{1}{g_{m1}R_s}\end{aligned} \tag{4.3}$$

3) Input referred noise due to R_D:

$$\frac{4kT}{R_D}\cdot\frac{R_D^2}{R_D^2+\omega^2L_D^2}\cdot\frac{Z_{load}^2}{Gain_{LNA}} = \frac{4kTR_D}{\left(R_D^2+\omega^2L_D^2\right)\cdot g_{m1}^2\cdot 4kTR_s\cdot\left[\frac{Z_{in}(s)}{Z_{in}(s)+R_s}\right]^2} \tag{4.4}$$

Summing up these three parts of noise contribution yields the total noise factor:

$$F = 1 + \frac{\gamma}{\alpha g_{m1} R_s} + \frac{\delta \alpha}{5 g_{m1} R_s}\left(\frac{\omega}{\omega_T}\right)^2 + \frac{R_D}{\left(\omega^2 L_D^2 + R_D^2\right) R_s g_{m1}^2 \left[\frac{Z_{in}(s)}{Z_{in}(s) + R_s}\right]^2} \tag{4.5}$$

Because L_c partially cancels the parasitic capacitance at the source node of cascode transistor M_2, its noise contribution remains much less than that of M_1 even at relatively high frequencies. The noise is dominated by the thermal noise (2nd term), which is mainly frequency-independent. The frequency-dependent gate induced noise (3rd term), and the frequency shaping of the resistor noise (4th term) result in a small variation of the CG-LNA noise factor over the BW.

4.3.3 Linearity analysis of the CG input stage

Fig. 4.5 shows the small-signal model for linearity analysis, where Z_{M1} is the impedance looking out of the drain of M_1.

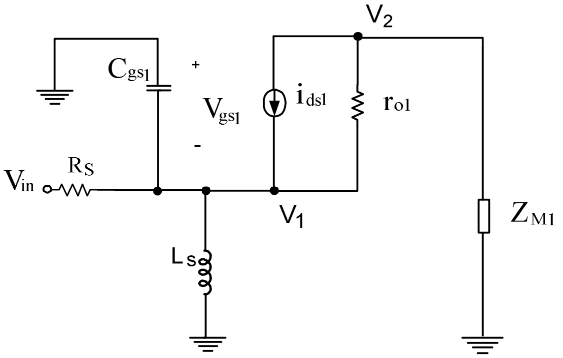


Fig. 4.5. Equivalent circuit of the CG-LNA input stage of Fig. 4.3 and Fig. 4.4.

The drain current of M_1 can be modelled up to 3rd order as:

$$i_{ds1} = -g_{m1}v_1 + g_2 v_1^2 - g_3 v_1^3 \tag{4.6}$$

where g_{m1}, g_2 and g_3 are the main transconductance and the 2nd/3rd order nonlinearity coefficients, respectively. Because capacitive and inductive (non-static) effects play an important role in LNA linearity, this work calculates the frequency-dependent harmonic-distortion coefficients using Volterra series (see Appendix B). The relation between the source voltage V_1, the drain voltage V_2, and the input voltage V_{in} can be expressed up to 3rd order as:

$$V_1 = A_1(\omega) \circ V_{in} + A_2(\omega_1, \omega_2) \circ V_{in}{}^2 + A_3(\omega_1, \omega_2, \omega_3) \circ V_{in}{}^3 \tag{4.7}$$

$$V_2 = C_1(\omega) \circ V_{in} + C_2(\omega_1, \omega_2) \circ V_{in}{}^2 + C_3(\omega_1, \omega_2, \omega_3) \circ V_{in}{}^3 \tag{4.8}$$

where "o" is the Volterra series operator, and $A_1(\omega)/C_1(\omega)$, $A_2(\omega_1, \omega_2)/C_2(\omega_1, \omega_2)$, and $A_3(\omega_1, \omega_2, \omega_3)/C_3(\omega_1, \omega_2, \omega_3)$ are the 1st-, 2nd-, and 3rd-order Volterra kernels [4]. $A_1(\omega)$ and $A_3(\omega_1, \omega_2, \omega_3)$ can be calculated as [see Appendix B.3 for detailed derivation]:

$$A_1(\omega) = \frac{r_{o1} + Z_{M1}}{H(\omega)} \tag{4.9}$$

$$A_3(\omega_1, \omega_2, \omega_3) = \frac{A_1^3 \cdot R_s \cdot r_{o1} \cdot \varepsilon\left(\Delta\omega, \omega_1 + \omega_2\right)}{H\left(\omega_1 + \omega_2 + \omega_3\right)} \tag{4.10}$$

$$H\left(\omega\right) = R_s + r_{o1} + Z_{M1} + g_{m1} r_{o1} R_s + R_s B\left(\omega\right)\left(r_{o1} + Z_{M1}\right) \tag{4.11}$$

$$\varepsilon(\Delta\omega, \omega_1 + \omega_2) = g_3 - g_{oB}(\Delta\omega, \omega_1 + \omega_2) \tag{4.12}$$

where $B(\omega) = j\omega C_{gs1} + 1/j\omega L_s$, $g_{oB}(\Delta\omega, \omega_1 + \omega_2) = \frac{2}{3} g_2{}^2 r_{o1} R_s \left[1/H(\Delta\omega) + 1/H(\omega_1 + \omega_2)\right]$, $\Delta\omega = \omega_1 - \omega_2$, and $H(\omega)$ represents a transimpedance relating the input voltage to the nonlinear drain current. $\varepsilon(\Delta\omega, \omega_1 + \omega_2)$ shows how the 2nd-order(g_{oB}) and 3rd-order(g_3) nonlinearity coefficients affect the 3rd order distortion. The capacitive effect at the source of M_1 is resonated out by the inductor L_s, thus $B(\omega)$ remains small over the BW. Therefore, under the input matching condition, $H(\omega)$ is simplified to a frequency-independent expression as:

$$H(\omega) = R_s + 2r_{o1} + Z_{M1} \tag{4.13}$$

The Volterra kernel in (4.8) is calculated as:

$$C_1(\omega) = -Z_{M1} \cdot A_1(\omega) \cdot \left(\frac{1}{R_s} + B(\omega) \right) + Z_{M1} \cdot \frac{1}{R_s} \tag{4.14}$$

$$C_3(\omega_1, \omega_2, \omega_3) = \frac{-Z_{M1} \cdot A_3(\omega_1, \omega_2, \omega_3) \cdot \left(1 + R_s B(\omega)\right)}{R_s} \tag{4.15}$$

A linear Z_{M1} results in a linear relation between $C_i(\omega)$ and $A_i(\omega)(i=1,3)$, and voltage V_2 is a linearly scaled version of V_1; however, if Z_{M1} is nonlinear, then V_2 is a distorted version of V_1. The expression for IIP3 can be written as [23]:

$$IIP3_{dBm} = 20 \cdot \log_{10} \left(\sqrt{\left| \frac{4}{3} \frac{C_1(\omega)}{C_3(\omega_1, \omega_2, \omega_3)} \right|} \right) + 10dB \tag{4.16}$$

$C_1(\omega)$ is usually fixed by the design parameters, therefore low distortion is achieved by reducing $C_3(\omega_1, \omega_2, \omega_3)$(i.e. by reducing $|\varepsilon(\Delta\omega, \omega_1 + \omega_2)|$). For a MOS transistor in saturation region, g_3 is negative and g_{oB} is positive, so simultaneously reducing g_3 and g_{oB} increases IIP3. The 2nd order feedback paths that contribute to 3rd order distortion in an LNA include the gate-drain capacitance C_{gd} [14] and the source degeneration inductor L_s [24], [25]. In the CG-LNA, the gate terminal of M_1 is AC grounded, reducing the feedback from C_{gd}. Thus, the 3rd-order distortion contributed by 2nd-order nonlinearity is smaller than in a CS-LNA. Section 4.4.3 compares these theoretical calculations to simulation results.

Under matching condition, the input impedance Z_{in} is estimated as $1/g_{m1}$, and (4.1) becomes:

$$G_m = \frac{g_{m1}}{1 + g_{m1} R_s} \tag{4.17}$$

Equation (4.17) is the same as for the resistive source degenerated transistor. Therefore, the linearity benefit of the resistive degeneration still holds true for the CG-LNA. On the other hand, the high Q series input matching network in the CS-LNA degrades its linearity because the Vgs is boosted by Q times. From the above discussion, the CG-LNA has a better linearity than the CS-LNA. For a typical design, a CG-LNA can achieve more than 5dBm IIP3 compared to the CS-LNA.

4.4 Proposed High Frequency Linearization Technique

4.4.1 Conceptual idea of the linearization method

The cascode LNA has slightly worse linearity than a single-transistor LNA due to the reduced headroom. Thus, the proposed linearization technique is implemented on the cascode LNA. Fig. 4.6 illustrates the conceptual idea. The additional transistor M_{1a} taps voltage V_2 and replicates the nonlinear drain current of the main transistor M_1, partially cancelling both the 2nd- and 3rd-order distortion terms.

The transistor-level implementation is shown in Fig. 4.7. The inductor L_c and the parasitic capacitances at the drain of M_1 and at the source of M_2 form a broadband π

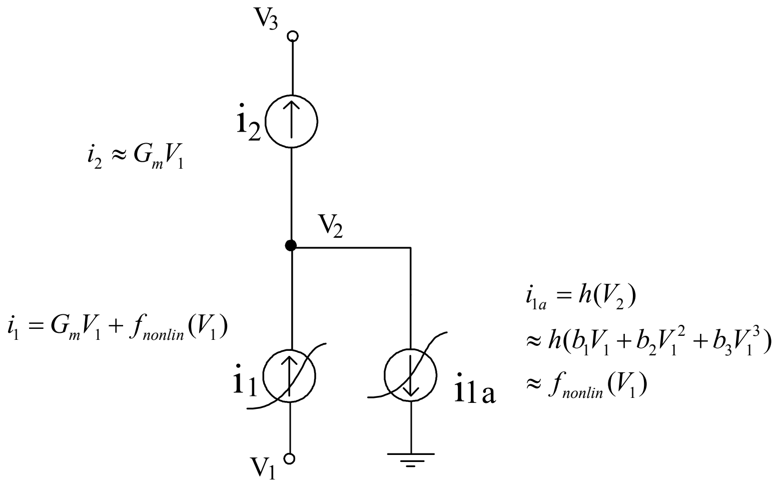

V3
i2
$i_2 \approx G_m V_1$
V2
$i_1 = G_m V_1 + f_{nonlin}(V_1)$
i1
i1a
$i_{1a} = h(V_2)$
$\approx h(b_1 V_1 + b_2 V_1^2 + b_3 V_1^3)$
$\approx f_{nonlin}(V_1)$
V1

Fig. 4.6. Conceptual idea of the linearization technique.

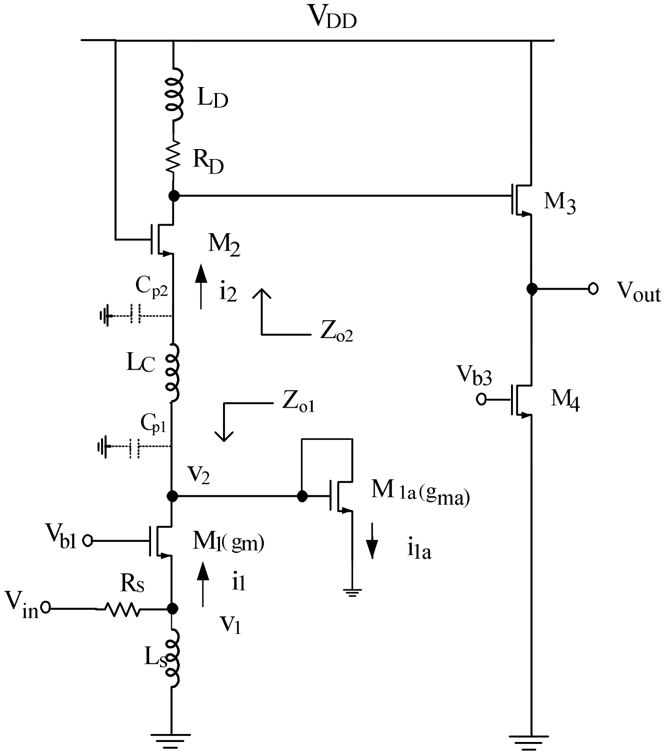


Fig. 4.7. Proposed linearized single-stage cascode UWB CG-LNA.

network. Proper choice of L_c cancels the capacitive effects, yielding effectively a short circuit over the whole BW. Under this condition, nonlinearity from M_2 can be neglected [15], leaving M_1 as the dominant source of nonlinearity, and Z_{o2} can be approximated as $1/g_{m2}$. Z_{o1} is the parallel combination of $1/g_{ma}$ and the output impedance of M_1. The diode connected transistor M_{1a} linearizes M_1 as follows. First, model the drain currents of M_1 and M_{1a} as:

$$i_1 = g_m v_1 + g_2 v_1^2 + g_3 v_1^3 \tag{4.18}$$

$$i_{1a} = g_{ma} v_2 + g_{2a} v_2^2 + g_{3a} v_2^3 \tag{4.19}$$

Next, suppose V_2 is related to V_1 by:

$$v_2 = b_1 v_1 + b_2 v_1^2 + b_3 v_1^3 \tag{4.20}$$

where b_1-b_3 are in general frequency dependent and can be extracted from simulation. In practice, the π network cancels the effects of b_2 and b_3 at the frequency of interest. The two nonlinear current i_1 and i_{1a} add up at node V_2, yielding the output current i_2:

$$i_2 = i_1 - i_{1a} = \left(g_m - b_1 g_{ma}\right) v_1 + \left(g_2 - b_1^2 g_{2a} - b_2 g_{ma}\right) v_1^2 + \left(g_3 - b_1^3 g_{3a} - g_{ma} b_3 - 2 g_{2a} b_1 b_2\right) v_1^3 \tag{4.21}$$

To obtain a good IIP3, the 3rd order distortion of the output current (3rd term in (4.21)) should be close to zero.

The output impedance at the drain of M_1, Z_{out}, is Z_{o1} // Z_{o2}. Z_{out}'s effect on linearity is twofold:

1) "2nd-order interaction" because of feedback: this has been addressed in Chapter II-Section 2.2.1 ("Feedback") and the solutions are provided in Section 2.2.2 ("Harmonic Termination") in this dissertation.
2) It modulates the drain current through V_{ds}. To model this effect, two-dimensional Taylor series can be used, which has been addressed in Chapter II-Section 2.4.

On the other hand, this proposed technique focus on linearizing the input transistor's transconductance, therefore we simplify the derivation by using parameters b_1-b_3 to model this "Z_{out} effect", as shown in equation (4.20).

The LNA is initially designed to satisfy input matching, gain, NF, and power. Next, M_{1a} is added to introduce additional degrees of freedom g_{ma}, g_{2a}, and g_{3a} to cancel the distortion from M_1. Without this auxiliary transistor, the only "knob" is the caocode transistor M_2, which affects the linearity indirectly through V_{ds} and is difficult to completely cancel the distortion. Though M_{1a} partially cancels the linear term as well, it does not appreciably degrade the gain/NF because its bias is much less than that of M_1. Finally, note that M_1 and M_{1a} uses identical finger sizes to improve matching and hence the cancellation of harmonics.

4.4.2 High-Frequency Analysis with Volterra Series

Fig. 4.8 shows the CG-LNA schematic for high frequency distortion analysis. Since the parasitic capacitance associated with the drain of M_1 is absorbed by the LC π network, it is not modelled here. The passive load resistance is much smaller than the transistor output resistance, thus we also neglect distortion due to nonlinear r_o. The analysis is limited up to 3rd order, assuming a weakly nonlinear circuit. Solving KCL equations together with equations (4.7) and (4.18)-(4.20), the 3rd order distortion of the output current ($i_{out,3rd}$) can be calculated as [see Appendix B.6 for detailed derivations]:

$$i_{out,3rd} = A_1(\omega)^3 \cdot \left[\begin{array}{l} \left(g_3 - b_1^2 g_{3a}\right) - \dfrac{g_3'\left(g_m - b_1 g_{ma}\right)}{H(\omega)} \\ + \dfrac{g_{oB}'\left(\Delta\omega, \omega_1 + \omega_2\right)\left(g_m - b_1 g_{ma}\right)}{H(\omega)} - \dfrac{2}{3}\dfrac{\left(g_2^2 - b_1^4 g_{2a}^2\right)}{H\left(\omega_1 + \omega_2\right)} \end{array} \right] \cdot v_{in}^3 \quad (4.22)$$

$$A_1(\omega) = \frac{1}{R_s} \cdot \frac{1 + r_{o1}/r_{o1a}}{H(\omega)} \quad (4.23)$$

$$H(\omega) = \left(\frac{1}{R_s} + B(\omega)\right)\left(1 + \frac{r_{o1}}{r_{o1a}}\right) + \frac{1}{r_{o1a}} + \frac{r_{o1}}{r_{o1a}} g_m + b_1 g_{ma} \quad (4.24)$$

where $g_3' = \dfrac{r_{o1}}{r_{o1a}} g_3 + b_1^3 g_{3a}$

$$g_{oB}'\left(\Delta\omega, \omega_1 + \omega_2\right) = \frac{2}{3}\left(\frac{r_{o1}}{r_{o1a}} g_2 + b_1^2 g_{2a}\right)^2 \left[\frac{1}{H(\Delta\omega)} + \frac{1}{H\left(\omega_1 + \omega_2\right)}\right],$$

$B(\omega) = j\omega C_{gs1} + 1/j\omega L_s$, ω_1 and ω_2 is the frequency of the two test tones.

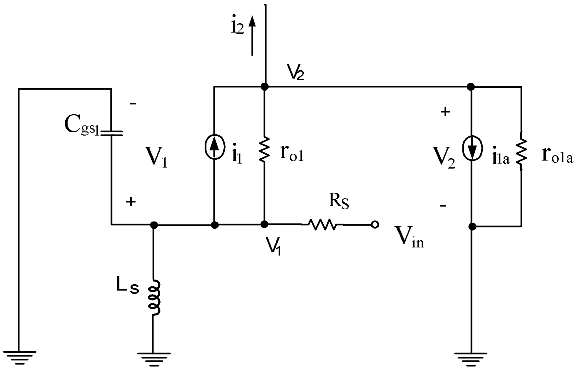


Fig. 4.8. Equivalent circuit for high frequency linearity analysis.

Equation (4.22) can be solved to obtain optimal IIP3. At the operating frequency, B(ω) ≈ 0, so H(ω) is a weak function of frequency. If $\omega_1 + \omega_2$ falls in band, then $H(\omega_1+\omega_2)$ is also only weakly frequency-dependent. If $\omega_1 + \omega_2$ is out of band, thanks to the low Q input matching network, the imaginary part in $H(\omega_1+\omega_2)$ is much smaller than the real part, making the frequency dependent effect still very small. Thus all the four terms in the bracket of (4.22) are approximately constant with respect to frequency, hence increasing the bandwidth of this linearization technique. This is verified by the measurement results shown later.

4.4.3 Comparison of Analytical Expressions and Simulations

Fig. 4.9 compares the IIP3 calculated with Volterra series to that computed in SPICE. Two -20dBm test tones separated by 100MHz were swept from 1~10GHz and applied to the cascode CG-LNA. As shown in Fig. 4.9, the theory predicts IIP3 frequency dependence quite well, the maximum deviation over the 1~10GHz band is less than 2dB. The obtained Volterra Series formulas (4.22)-(4.24) can also predict the IIP3 variation as a function of two-tone spacing, as will be presented in Section 4.5.

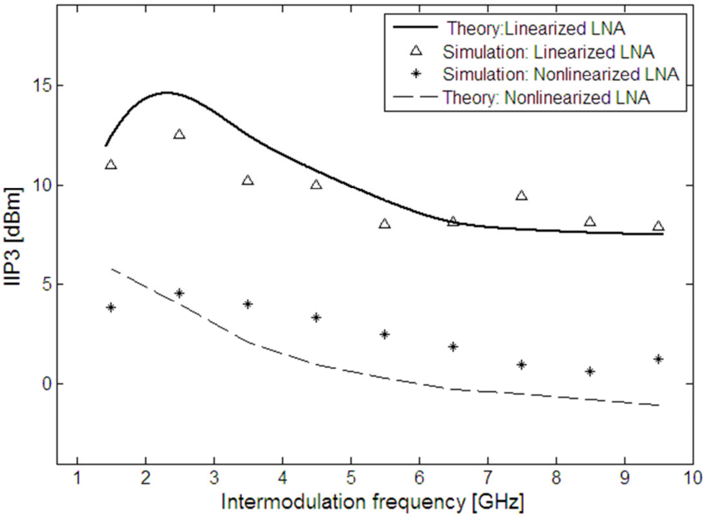


Fig. 4.9. IIP3 comparison of analytical expressions (4.13) and (4.19) with SPICE simulations for cascode LNA with and w/o linearization, using 100MHz spacing two-tone with -20dBm power level.

4.4.4 Process and Temperature Variations

To investigate the temperature sensitivity of the proposed linearization technique, post-layout IIP2 /IIP3 simulations were conducted at -40°C, 27°C, and 85°C. IIP2 tests fixed one tone at 2.4GHz and the other tone at 5.4, 3.1, and 5.6GHz. IIP3 tests used 30MHz tone spacing. In all cases, Pin = -20dBm. The 3GHz, 5GHz, and 8GHz in Table 4.2 mean the intermodulation frequency (IM2 or IM3). IIP3 and IIP2 improvement above 4.4dB and 4.7dB respectively are achieved across temperature. The main effect of temperature variations is on $g_m(T)$ and $g_{ma}(T)$, since both M_1 and M_{1a} work in saturation region and have same unit finger size, good matching is guaranteed, hence robust distortion cancellation is maintained across temperature variation.

Table 4.2. Linearity improvement versus temperature
(Post-layout simulation with two input tones at -20dBm)

	Temperature Inter-modulation frequency	-40°C	27°C	85°C
IIP3 Improvement (dB)	3GHz	11.5	8.2	6.1
	5GHz	8.2	7.8	6
	8GHz	9.1	5.8	4.4
IIP2 Improvement (dB)	3GHz	10.2	8.5	4.7
	5GHz	14.7	13	10.6
	8GHz	9.4	7	5.3

To check the effect of process variation, pre-layout simulation was performed with a ± 20% variation in the size of M_{1a}. Consistent IIP3 and IIP2 improvement above 7dB and 5dB respectively is obtained over the BW. These results verify the effectiveness of the linearization technique in a wide frequency range across process and temperature variations.

4.5 Effects of the Linearization Technique on S11, and NF

Because $g_{m,M1a} << g_{m,M2}$, the input impedance $Z_{in}(s)$ seen from R_s of the CG-LNA is about the same with and without linearizing transistor M_{1a} present. Thus, M_{1a} does not significantly affect matching. This is confirmed in both the simulation and measurement.

The small-signal model for noise analysis is shown in Fig. 4.10. Besides the input referred noise due to the channel noise and gate noise from M_1, and the thermal noise from load resistor R_D, we also need to add the channel noise and gate noise from M_{1a} to obtain the total noise factor.

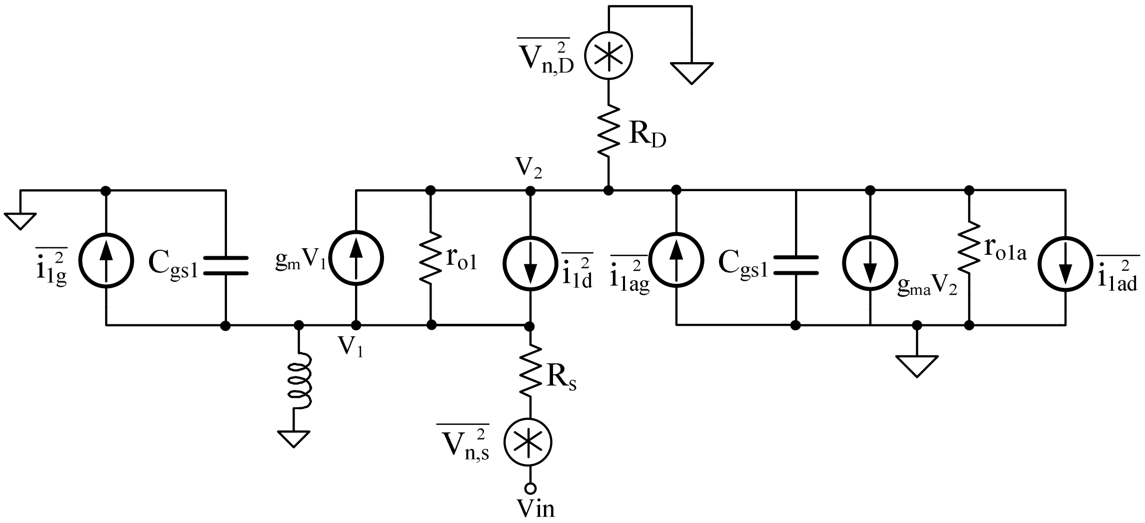


Fig. 4.10. Small-signal model for noise analysis of the linearized cascode CG-LNA.

The channel noise and gate-induced noise of M_{1a} appearing at the LNA output is:

$$\overline{i^2{}_{nd,M_{1a}}} = 4kT\frac{\gamma}{\alpha}g_{m1a} \tag{4.25}$$

$$\overline{i^2{}_{ng,M_{1a}}} = 4kT\frac{\delta\alpha\omega^2 C^2{}_{gs,1a}}{5}\cdot\frac{g_{m1a}}{\left(g_{m1a}+g_{m2}\right)^2} \tag{4.26}$$

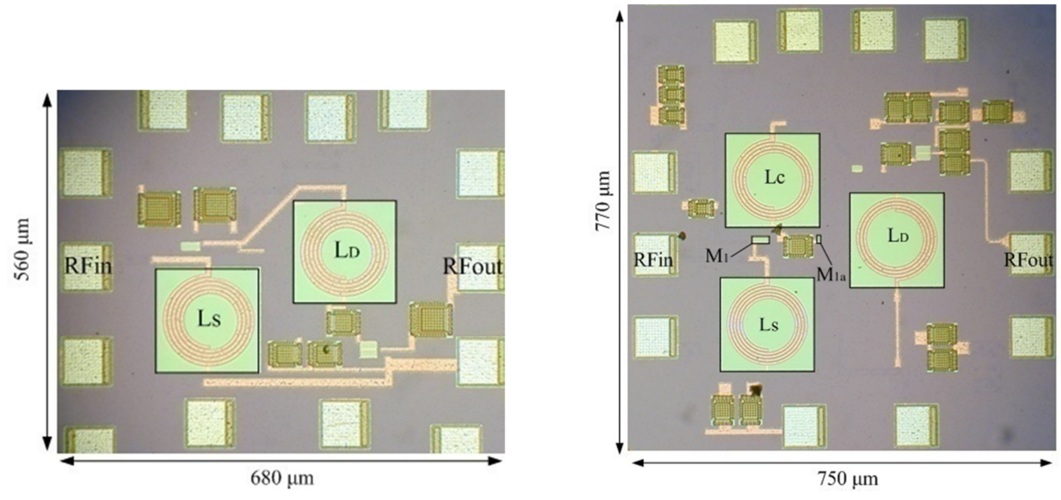
The noise contribution from M_{1a} is proportional to its transconductance(i.e. g_{m1a}), which is much smaller than g_{m1}. The noise factor of the proposed linearized cascode CG-LNA can be calculated as:

$$F = 1 + \frac{\gamma}{\alpha g_{m1} R_s} + \frac{\delta\alpha}{5 g_{m1} R_s}\left(\frac{\omega}{\omega_T}\right)^2 + \frac{R_D}{\left(\omega^2 L_D^2 + R_D^2\right) R_s g_{m1}^2 \left(\frac{Z_{in}(s)}{Z_{in}(s) + R_s}\right)^2}$$
$$+ \frac{\gamma}{\alpha g_{m1} R_s} \cdot \frac{g_{m1a}}{g_{m1}} + \frac{\delta\alpha}{5 g_{m1} R_s}\left(\frac{\omega}{\omega_T}\right)^2 \cdot \frac{g_{m1a}}{g_{m1}} \cdot \frac{g^2{}_{m1a}}{\left(g_{m1a} + g_{m2}\right)^2} \qquad (4.27)$$

In (4.27), the last two terms are the additional noise contribution from the linearization circuitry, while the first four terms are from the cascode CG-LNA without linearization, as shown in (4.25). The 5th term is the channel noise of M_{1a}, which is smaller than the channel noise of M_1 by a factor of g_{m1a}/g_{m1} (0.07 in our design). The 6th term-- M_{1a} gate induced noise-- is $(g_{m1a}/g_{m1})^3$ = 3.4e-4 times smaller than the gate noise in M_1. Thus the degradation in NF is small--less than 0.6dB over the entire measured BW. Based on the above discussion, the proposed linearization technique does not appreciably affect the input matching and NF.

4.6 Measurement Results

Both a single-transistor and a cascode single-stage UWB CG-LNA were fabricated in UMC 0.13μm CMOS technology. The proposed linearization technique is implemented on the cascode LNA. The chip micrograph is shown in Fig. 4.11. The single-transistor CG-LNA core occupies 320μm×420μm, and the cascode CG-LNA core uses 480μm×480μm. The output buffer effect is de-embedded from the LNA+Buffer measurements using the measured results of a fabricated stand-alone buffer.



(a) (b)

Fig. 4.11. Chip Micrograph of a) single-transistor UWB CG-LNA b) cascode UWB CG-LNA.

On-wafer probing was performed to measure these UWB LNAs. Fig. 4.12 shows the lab measurement setup. Two single-ended RF probes were used to feed the input signal and take the output signal. The DC probes were used to provide DC bias and power supply. On-wafer probing facilitates this high frequency testing because the bond wire effects and PCB parasitic were eliminated. The disadvantages are: 1) hard to use commercial regulators to filter out the power supply noise; 2) the available DC probes are limited, thus cannot provide many DC bias.

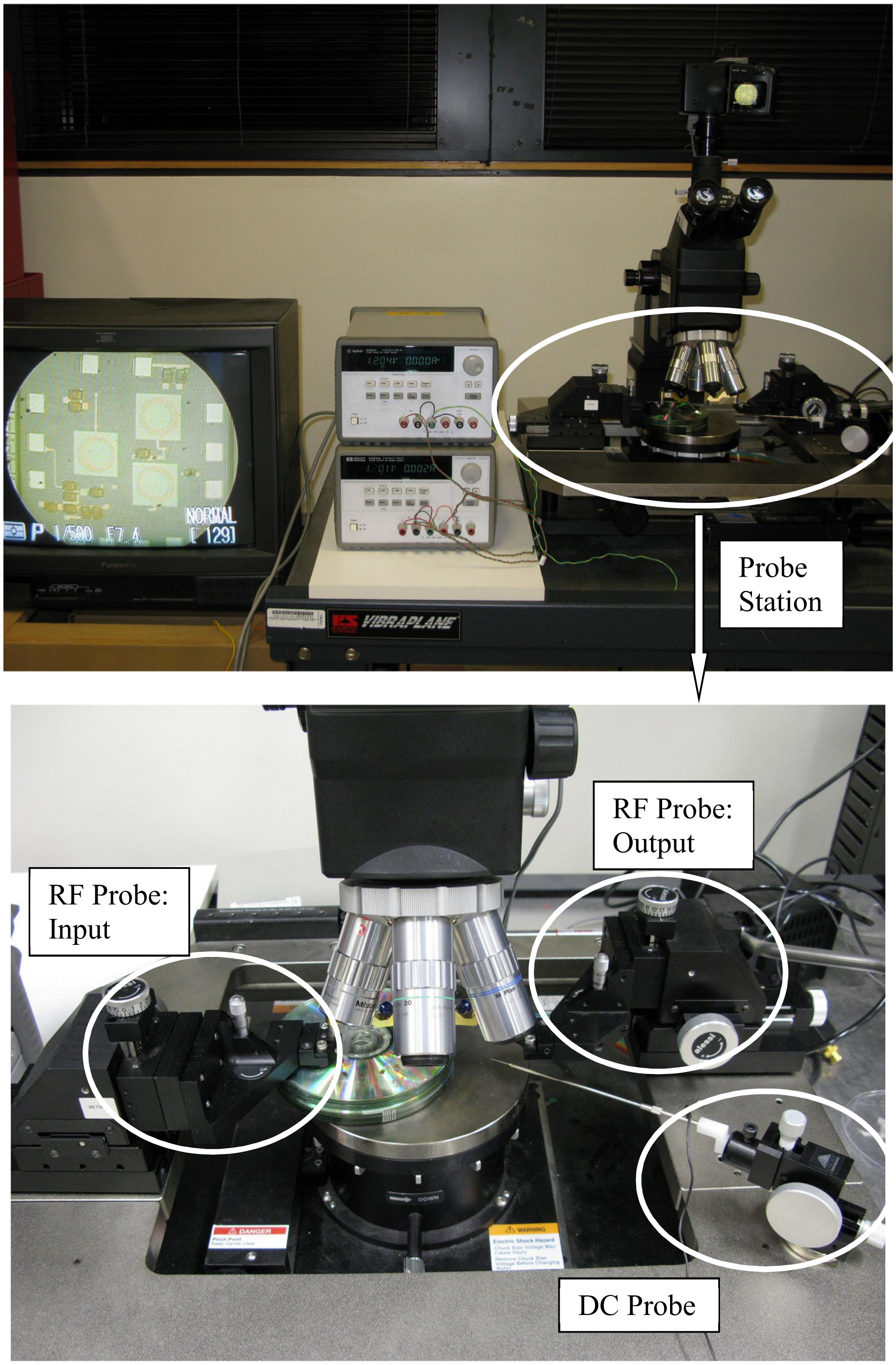


Fig. 4.12. On wafer probing lab measurement setup.

4.6.1 Single-Transistor CG-LNA

Fig. 4.13 shows a maximum measured gain S_{21} = 10dB with max variation $\underline{+}$1.5dB over 3-11GHz BW. S_{11} < -10dB at high frequency (up to 12GHz) but degrades slightly around 3GHz. Fig. 4.14 indicates a minimum NF = 2.9dB and variation < 0.7dB over 3-10GHz.

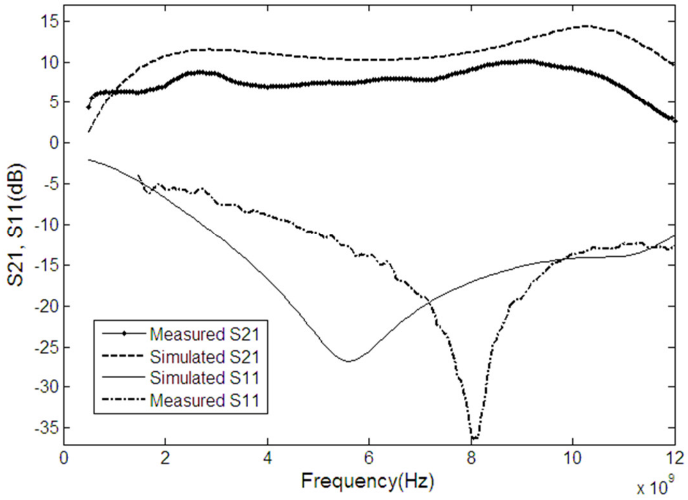


Fig. 4.13. S11 and S21 of the single-stage single-transistor CG-LNA (Fig. 4.3).

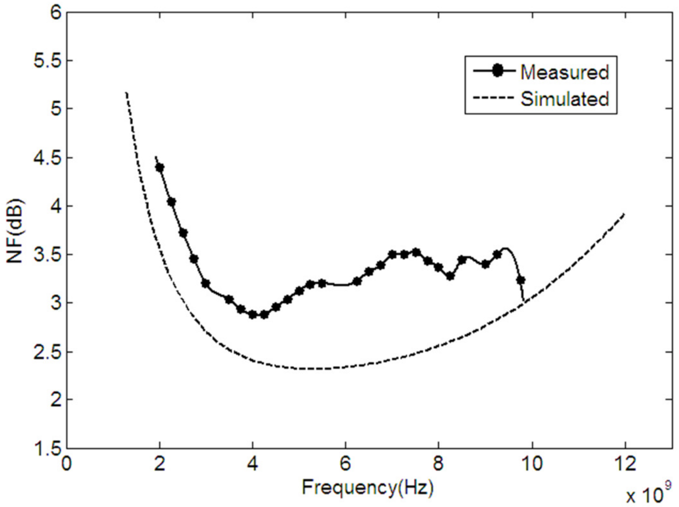


Fig. 4.14. Measured & simulated NF of the single-transistor CG-LNA.

In wideband operation, widely spaced tones will in practice dominate the IIP3 and IIP2. For example, the potential interferers for the UWB system include GPS, PCS/DCS, UMTS, ISM band (802.11b/g, Bluetooth, Zigbee, IEEE 802.15.2, Microwave ovens), WiMax, and IEEE 802.11a. Thus the intermodulation products from the interferences with frequency spacing between tens of MHz to GHz need to be considered. For IIP2 measurement, one input tone is fixed at 5.2GHz, while the other changes from 3GHz to 9GHz. For IIP3 measurement, we use two tones with 30MHz spacing at: 2.8GHz, 4.1GHz, 5.1GHz, 6.1GHz, 7.1GHz, 8.1GHz, 9.1GHz, 10.1GHz, and 11.1GHz. Fig. 4.15 shows the measured IIP2/3 performance of the single-transistor LNA. IIP2/3 were computed by IIP2 = $2P_{in} – P_{IM2}$ and IIP3 = $P_{in} + 0.5^* (P_{in} – P_{IM3})$, respectively, where P_{in} is the input power of one test tone in two-tone test, and P_{IMk} indicates the input referred power of the k^{th} order intermodulation tone. In all cases, input tones have P_{in} = -20dBm. As shown in the figure, the single-transistor LNA achieves an IIP2 of 5-15dBm and an IIP3 of 6.5-9.5dBm. The measured IIP3 versus frequency spacing (Δf_{in}) for the cascode LNA will be shown later. The UWB single-transistor LNA consumes only 1.85mA from a 1.3V power supply.

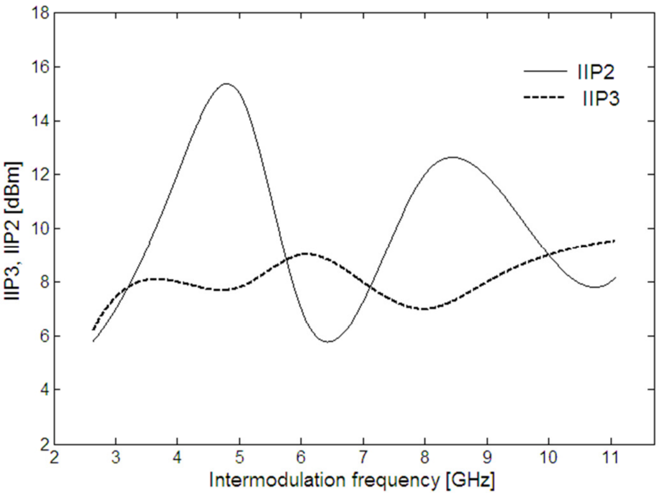


Fig. 4.15. Experimental IIP2, IIP3 for single-transistor LNA at different input frequencies.

4.6.2 Cascode CG-LNA

Fig. 4.16 shows an S_{11} < -10dB over the 2.7GHz-12GHz frequency range. As predicted by theory in Section V, the linearization method hardly affects S_{11}. The discrepancy between simulation and test results is mainly attributed to the extra parasitic effects, and is significantly reduced when a 70fF extra capacitance is added to the input node in simulation. As shown in Fig. 4.17, a 12.6dB maximum gain with ±1.5dB variation is obtained over the 0.8GHz-8.4GHz BW before linearization, the gain degradation remains below 1.7 dB over the entire band after linearization. As shown in Fig. 4.18, the LNA has a minimum NF as 3.3dB, and a ±0.75dB variation before linearization; the degradation in NF is less than 0.6 dB over the entire band after linearization. The variations in the NF arise from the frequency-dependent gate induced noise and the load resistor noise as shown in (4.5). The cascode transistor also contributes some frequency-dependant noise [15]. The UWB cascode LNA consumes only 2mA from a 1.3V power supply, and the linearization element only draws an additional 20μA.

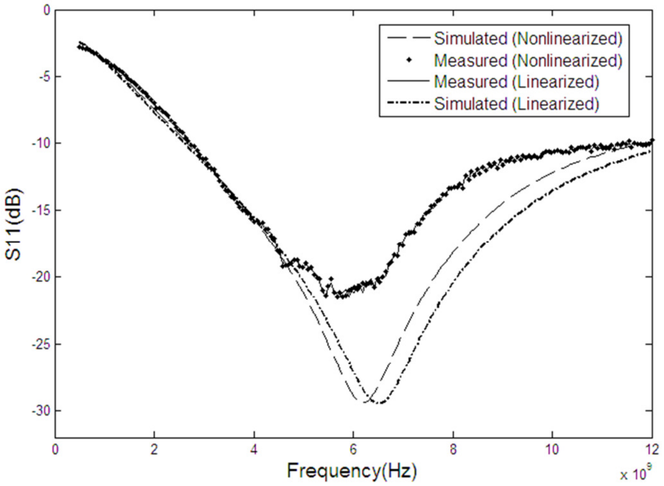


Fig. 4.16. Measured & Simulated S11 of the Cascode CG-LNA with (Fig. 4.7) and w/o linearization (Fig. 4.4).

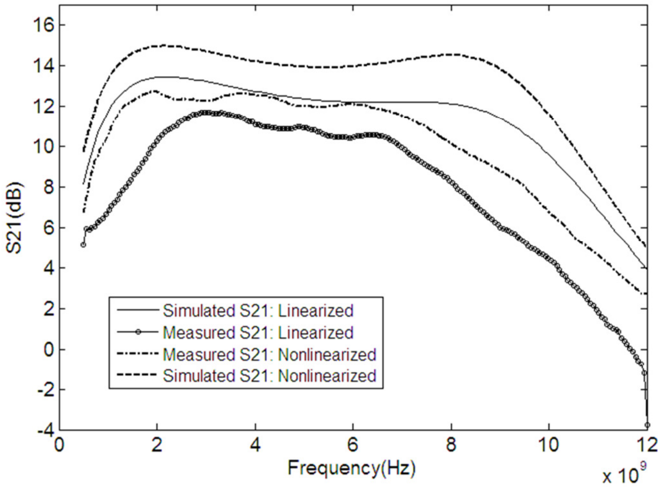


Fig. 4.17. Measured & simulated S21 of the cascode CG-LNA with (Fig. 4.7) and w/o linearization (Fig. 4.4).

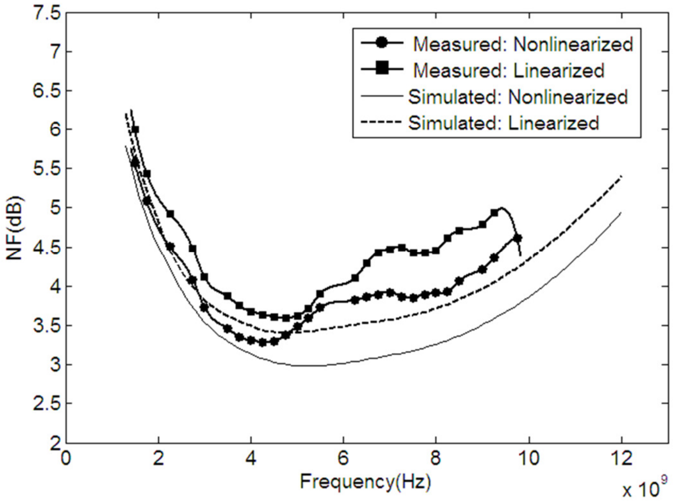


Fig. 4.18. Measured & simulated NF of the cascode CG-LNA with (Fig. 4.7) and w/o linearization (Fig. 4.4).

4.6.3 Design Robustness

To experimentally verify the robustness of the linearization technique, the IIP3 and IIP2 of the cascode LNA with/without linearization were measured on ten randomly chosen chips. The IIP3 of the cascode LNA was examined at seven different frequencies with 30MHz frequency spacing at -20dBm: 2.8GHz, 4.1GHz 5.1GHz, 6.1GHz, 8.1GHz, 9.1GHz, and 10.1GHz. As shown in Fig. 4.19, an IIP3 improvement greater than 3.5dB is achieved in worst case, while other samples showed an improvement as high as 9dB. For IIP2 measurement, one input tone is fixed at 5.2GHz, while the other changes from 3GHz to 9GHz, with equal power level as -20dBm. Fig. 4.20 shows IIP2 improvement > 3.3dB in the worst case and improvement up to 10dB in the best case. These results confirm the effectiveness and robustness of the proposed linearization technique over a wide frequency range. Because our technique utilizes all transistors in the saturation region, we obtain much better matching than previously reported methods that mixed and matched triode/weak-inversion transistors [14][19] [24][25].

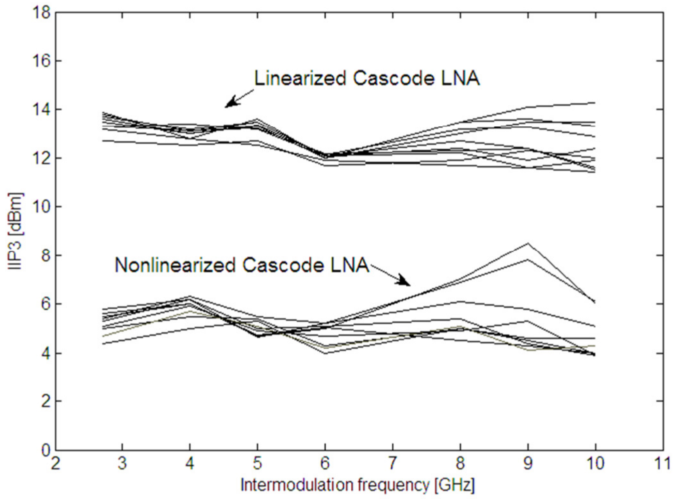


Fig. 4.19. Measured IIP3 (Cascode LNA) vs. intermodulation frequency (10 samples).

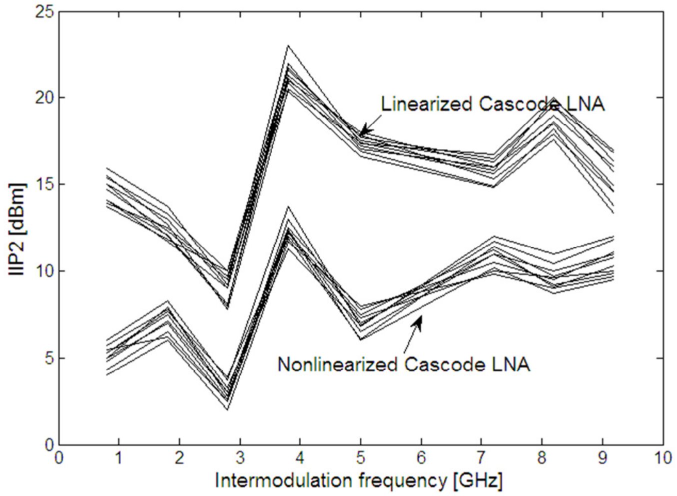


Fig. 4.20. Measured IIP2 (Cascode LNA) vs. intermodulation frequency (10 samples).

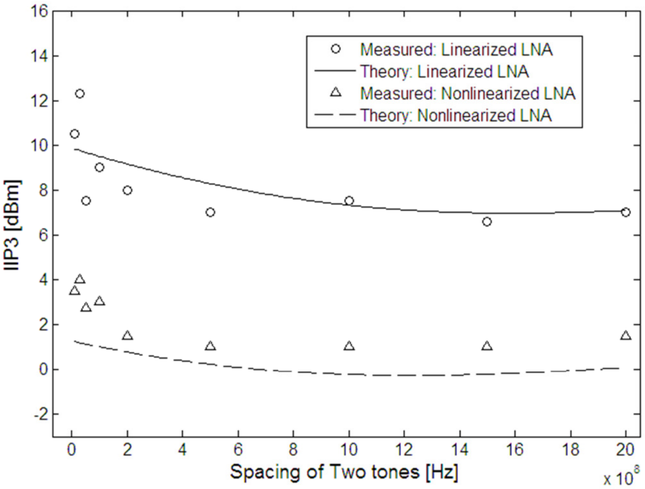


Fig. 4.21. Experimental and theoretical results of IIP3 for cascode LNA with and w/o linearization, as a function of frequency spacing for -20dB input tones.

To check the sensitivity of IIP3 to two-tone spacing (Δf_{in}), IIP3 was also measured by fixing one input tone at 5GHz while changing the other from 5.01GHz to 7GHz. Fig. 4.21 shows experimental and theoretical results (from (4.16) and (4.22)) of the IIP3 as a function of Δf_{in}. IIP3 degrades by 4dB when Δf_{in} increases from 10MHz to 200MHz, and stays relatively constant with a variation less than 1dB when Δf_{in} increases up to 2GHz. The Volterra series analysis in (4.7)-(4.16), (4.22)-(4.24) also indicates this trend. When Δf_{in} is small, the parallel tank formed by L_s and C_{gs1} at the input has large susceptance (i.e. $B(\Delta\omega)$ is large), resulting in larger $H(\Delta\omega)$, smaller g_{oB}, and hence a smaller 3rd distortion coefficient. As Δf_{in} increases, $B(\Delta\omega)$ decreases and remains small over the BW because of the low Q resonant network.

4.6.4 Gain and Linearity

These two LNAs' gains are low because 1) they have only one stage; 2) their g_ms are limited by input matching; and 3) their output impedance is low(due to headroom limitations). The IIP3 is high not because of low gain because the primary source of nonlinearity is the drain current generated by the input transistor--hence a high impedance load will not degrade linearity provided that it is linear and does not disturb the transistor bias points. To prove this, a simulation is conducted: keeping the bias of input transistor constant, thus g_m, g_{ds}, and I_{ds} is constant; change the load resistor R_D and the power supply accordingly to keep the drain-to-source voltage of transistors constant. The inductor L_D is also adjusted to maintain a flat gain over the BW. In this way we can vary the gain of the LNA to see its effect on IIP3. Two tones of 3.5GHz and 3.65GHz with -20dBm power are used. Shown in Fig. 4.22, the IIP3 of the cascode LNA without linearization degrades 3.15dB when gain varies from 6.8dB to 15.8dB; but an IIP3 improvement of 3.5~6.5dB can be obtained over the whole gain variation range by applying the linearization technique. The small variation of 2dB in IIP3 of the linearized cascode LNA with increasing gain proves that high IIP3 is not due to low gain.

The simulation result also proves that the proposed linearization technique is effective no matter what the LNA gain is. Thus as a general linearization technique, it can be applied to other LNA topologies, either with high gain or low gain. The only condition is, the linearization element must be added to a low impedance node in order not to load the original LNA. This simulation also demonstrates that the LNA has the potential of obtaining larger gain thus better NF, while maintaining excellent linearity.

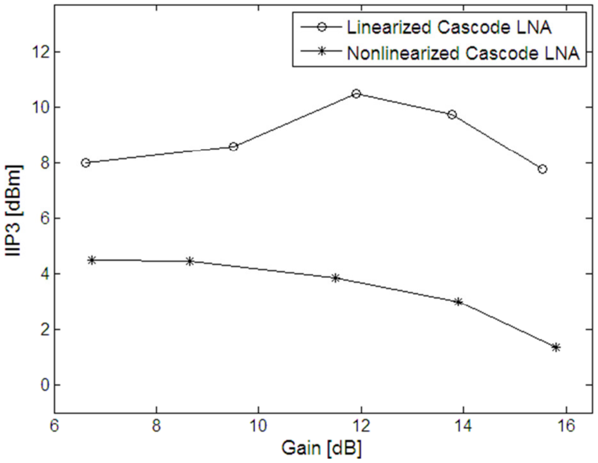


Fig. 4.22. Simulated IIP3 vs. gain.

This retains the advantage of low power consumption, at the cost of larger area and supply voltage. But in many applications the RF power amplifier and baseband analog signal processing circuits also run from a higher supply voltage than 1.2 V [29], [78], making this a viable alternative.

In the UWB "impulse radio" application, linear phase response across BW is also required to minimize phase distortion and recover the transmitted signal correctly. The S21 phase versus frequency for cascode LNA with and without linearization is simulated, and the maximum group delay variation is < 14% over the entire BW. The linearization technique adds negligible group delay deviations.

Experimental results of the proposed LNAs and the prior published state-of-art UWB LNAs are summarized in Table 4.3, in which the best data per column is marked with gray color. For the comparison of different topologies, we include two figures of merit (FOMs) in the table—FOM_I [75], which does not include linearity, and FOM_II [79], which does:

$$FOM\_I = \frac{Gain_{average}\left[abs\right] \times BW\left[GHz\right]}{P_{dc}\left[mW\right] \times \left(F_{average} - 1\right)} \tag{4.28}$$

$$FOM\_II = \frac{IIP3_{average}\left[mW\right] \times Gain_{average}\left[abs\right] \times BW\left[GHz\right]}{P_{dc}\left[mW\right] \times \left(F_{average} - 1\right)} \tag{4.29}$$

where $\mathrm{Gain}_{\mathrm{average}}$ is the average gain, $\mathrm{F}_{\mathrm{average}}$ is the average noise factor over the frequency range, and P_{dc} is the power consumption of the LNA core. From Table 4.3, our proposed LNAs achieve comparable IIP3 with much less power than the previously reported best linearity in [75] and [82]. This is mainly due to the simple input matching network, single-stage architecture, and the proposed linearization technique. All three proposed LNAs exhibit comparable FOM_I, and much better FOM_II when compared to the other state-of-art UWB LNAs. The FOM_II of the linearized cascode LNA exhibits a factor of 2.4 over the best previously reported result in [82].

4.7 Conclusions

In this chapter, a practical linearization technique is proposed for a UWB LNA, and a detailed linearity analysis using Volterra series is provided, which shows good agreement with simulation and experimental results. Three low-power single-stage UWB CG-LNAs are presented in this paper, with focus on the cascode LNA with linearization. The linearity of the proposed LNAs without linearization is also good because of the CG and single-stage topology. The UWB LNA was designed and fabricated in 0.13μm UMC CMOS technology. Because all transistors operate in the saturation region, we obtain a robust linearity improvement over process and temperature variations. The proposed linearization method is experimentally demonstrated to improve the IIP3 by 3.5 to 9dB over a 2.5~10GHz frequency range. A comparison of measurement results with the prior published state-of-art UWB LNAs shows that our proposed linearized UWB LNA achieves excellent linearity with much less power than previously published works.

Table 4.3. Measurement results summary and CMOS UWB LNA comparison

Parameters	3 dB BW [GHz]	S11 [dB]	S21 [dB]	NF [dB]	IIP3 [dBm]	IIP2 [dBm]	Power (core) [mW]	Area [mm^2]	FOM I	FOM II	CMOS Process
ST CGLNA	3~11	<-7.5	7~10	2.9~3.6	6.5~9.5	5~15	2.4	0.38	22.3	149	0.13μm
Cascode CGLNA	0.8~8.4	<-9	9.6~12.6	3.3~5.5	3.9~8.5 [5]	1.8~13.9 [5]	2.6	0.58	21.8	109	0.13μm
Linearized CGLNA	1.5~8.1	<-9	8.6~11.7	3.6~6	11.7~14.1 [5]	7.6~23 [5]	2.62	0.58	12.9	261.6	0.13μm
[70]STD LNA	2.3~9.2	<-9.9	6.3~9.3	4~9	-6.7 [2]	-3~7	9	1.1	1.2	0.2	0.18μm
[70]TW LNA	2.4~9.5	<-9.4	7.4~10.4	4.2~9	-8.8 [2]	-10~0	9	1.1	1.5	0.2	0.18μm
[73]	0.4~10	<-10	9~12.4	4.4~6.5	-6 [2]	-	12	0.42	3.9	0.97	0.18μm
[74]	1.2~11.9	<-11	6.7~9.7	4.5~5.1	-4.9~-6.2	9.8~20	20	0.59	1.85	0.52	0.18μm
[30]	0.2~5.2	<-10	13~15.6	2.9~3.5	0~4	18~34 [5]	14	0.009	9.22	16.22	65nm
[75] LNA #1 [1]	1.3~10.7	<-6	6.1~8.5	4.4~5.3	7.4~8.3	-	4.5	1.0	5.63	34.5	0.18μm
[75] LNA #2 [1]	1.3~12.3	<-6	5.2~8.2	4.6~5.5	7.6~9.1	-	4.5	1.0	7.4	51.4	0.18μm
[78]	3.1~10.6	<-9.9	13.7~16.5	2.1~2.9	-8.5~-5.1	-	9	0.87	35.97	8.1	0.13μm
[80]	0.1~6.5	<-10	17~19	3~4.2	+1^{+}	+4 [4]	12	-	26.37	33.23	0.13μm
[81]	2~11	<-8	9~12	5~6	-4 [3]	-	16.8	0.7	2.1	0.9	90nm
[82]	2~8	-	12	2.5	+12	-	18	-	6.77	107.4	65nm
[83]	0.6~10	-	10	3	+6	-	30 [6]	-	-	-	45nm

[1] Differential LNA [2] at 6GHz [3] at 4GHz [4] simulated [5] for multiple samples

[6] including VtoI converter

ST LNA: Single-Transistor LNA

CHAPTER V

NEW TRENDS IN ADC DESIGN*$^{+}$

5.1 Introduction

There are two salient factors that continuously motivate research activities on A/D converters, new applications and technology scaling. Emerging adaptive systems such as software defined radios and multi-sensor systems require an A/D interface integrated with multiple specifications and more intelligent energy utilization. Conventional ADCs are designed with fixed speed/resolution, the same large power is wasted when conditions become more favorable; an adaptive ADC reconfigures its speed/resolution based on instantaneous conditions, thus these schemes dramatically improve average power consumptions and extend the battery life [84]-[101].

A popular FOM for ADC [133] is:

$$FOM = \frac{Power}{2^{ENOB} \cdot f_s} \tag{5.1}$$

*©[2009] IEEE. Reprinted, with permission, from "New Applications and Technology Scaling Driving Next Generation A/D Converters", by Heng Zhang, Mohamed Mostafa Elsayed, and Edgar Sánchez-Sinencio, *IEEE European Conference on Circuit Theory and Design(ECCTD),* Antalya, Turkey, Aug.23-27 2009, pp.109-112.

$^{+}$©[2009] IEEE. Reprinted, with permission, from "Minimum Current/Area Implementation of Cyclic ADC", by Heng Zhang, Qunying Li, and Edgar Sánchez-Sinencio, *IEE Electronics Letters*, Vol. 45 Issue 7, pp.351-352, March 2009.

where the power includes both the analog and digital power of the ADC, f_s is the ADC sampling rate, and ENOB is the effective number of bits. This FOM has the unit of "J/Conversion-step", i.e. how much energy is consumed for one conversion step; therefore, it well quantifies the ADC efficiency. Some new applications with stringent energy constraint require ultra-low power ADCs. These systems include wireless sensor networks, implantable medical devices, and built-in testing. The key idea to increase ADC power efficiency is to eliminate the power-hungry opamps [104]-[109], and use digital circuitries to compensate the increased non-idealities due to simplified analog blocks [103], [110].

A recent trend in ADC design that leverages the strength of ultra-deep-submicron technologies, is time-domain-based ADCs [115], [116] in which the information is represented by a pulse width then a time-to-digital converter is used to get the digital output data. The increasing time-resolution associated with technology scaling makes this approach very attractive for the next generation ADCs.

Fig. 5.1 illustrates the two driving forces and three new trends for next generation ADCs.

5.2 Adaptive ADCs

5.2.1 Motivations

Nowadays we are expecting a migration from existing multi-standard transceiver to the promising "software defined radio". The reconfigurable ADCs for software defined radio should cover a larger range of specifications with better power scalability.

The ADCs for some popular communication standards should provide resolution from 4 to 14bits, while covering the signal bandwidth from 200k to 500MHz [84].

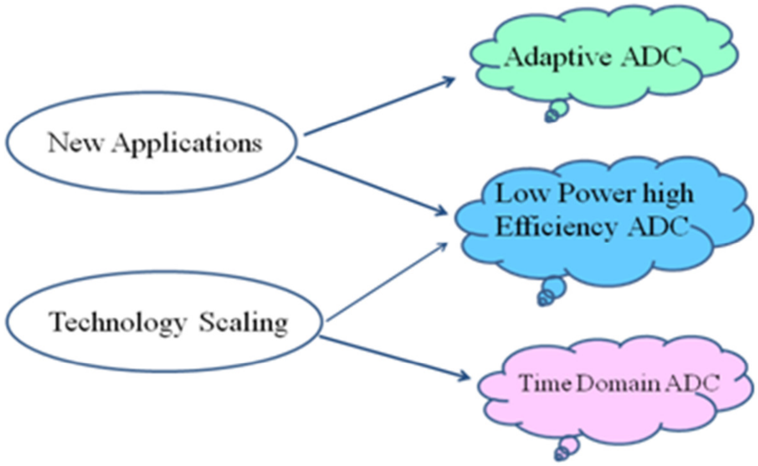


Fig. 5.1. Two driving forces and three new trends for next generation ADCs.

In another scenario, such as a multi-sensor system [86], the reconfigurable ADC should have multi-signal conversion capability to handle a variety of different signals (eg. voice, sound, image, temperature, seismic, blood pressure, heart beat, etc.) in real time.

An adaptive ADC differentiates from a reconfigurable ADC in terms of intelligence level. Fig. 5.2 shows the conceptual diagram of an adaptive ADC. The core part is a resolution/speed/power reconfigurable ADC. Two control paths are added. The first one is a dynamic controller, which senses the input signal information (rms power, bw, etc.) and send the reconfigure command to ADC. The other path is a digital controller for user interface; it allows users to program the resolution & speed of ADC according to their needs. An adaptive ADC with more intelligence can better fit the future communication systems and universal sensors than a reconfigurable ADC.

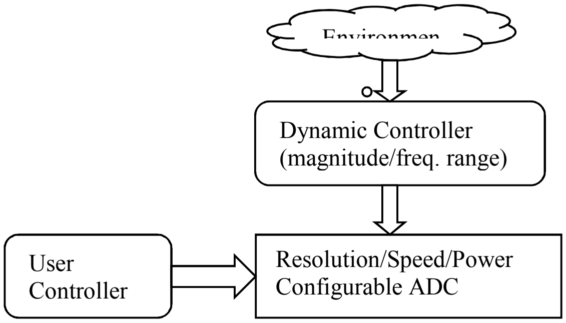


Fig. 5.2. Conceptual diagram of an adaptive ADC.

5.2.2 Topology Comparison

Table 5.1 summarizes the state-of-art silicon-verified reconfigurable ADCs. Pipeline and sigma-delta based architecture are most widely used. Sigma-Delta ADC allows an easy trade-off between sampling rate and dynamic range, thus is suitable for cellular applications. On the other hand, pipeline ADC has inherent higher speed, while its resolution envelope has been continuously expanded by digital calibration techniques. Fig. 5.3 shows the ADC performance from ISSCC 2008-2010 for pipeline, sigma-delta, SAR, and flash ADCs. It is observed that pipeline ADC is breaking the trend set by sigma delta, SAR, and flash ADCs, and is expected to be a key ADC architecture in future applications.

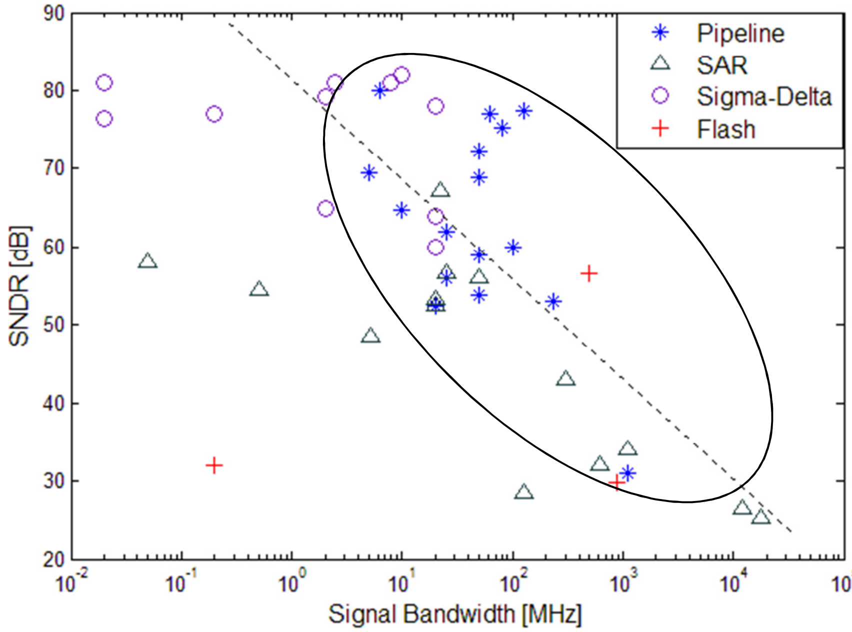


Fig. 5.3. Performance comparison of pipeline, sigma-delta, SAR, and flash ADCs (ISSCC 2008-2010).

5.2.3 Pipeline ADC Reconfiguration Methodology

A reconfigurable ADC can be viewed as an ADC with a configurable switch matrix. The switches should be maximally reused to reduce performance degradation. Pipeline ADC configurable parameters include the number of stages, resolution per stage, sampling rate, and number of time- interleaving branches.

The speed of pipeline ADC can be programmed in either architecture or circuit level. Since pipeline ADC share the same building blocks as cyclic ADC [102], its effective speed can be reduced by configuring into cyclic mode [96]. On the other hand, speed can be scaled by changing the biasing current of the active building blocks (i.e. Opamps). The resolution can be programmed by shortening the pipeline [86].

Table 5.1. State-of-art silicon-verified reconfigurable ADCs

	Architecture	Resolution (Bits)	Speed (MSPS)	Power (mW)	Process
[87]	Flash	4	10~1000	9.6~10.6	0.18µm
[88]	SAR	9	0~50	0~0.7	90nm
[89]	SAR	8/12	0-0.1/0-0.2	0~0.025	0.18µm
[90]	Pipeline	12	20~130	21~110	0.18µm
[91]	Pipeline	10	0.11~50	0.015~35	0.18µm
[92]	Pipeline	14	10/20/30/40	19/34/51/73	0.18µm
[93]	Pipeline	10	25~120	10~36	90nm
[94]	Time-interleave Pipeline	5/7	550/1100	13/30/46	90nm
[95]	Time interleave Pipeline	10/11	11/44	14.8/20.2	0.25µm BiCMOS
[96]	Hybrid Pipeline/Cyclic	6~10	2.5/5/10/80	30.2~93.7	0.18µm
[86]	Hybrid Pipeline/Sigma-Delta	6~16	0~10	2~24.6	0.6µm
[85]	Hybrid Pipeline/ Sigma-Delta	10/11/12	0.2/4/20 *	15/37	0.18µm
[97]	Sigma-Delta	11~15	0.1~10 *	2.9~20.5	0.13µm
[98]	Sigma-Delta	9~14	0.2~10 *	1.44~7	90nm
[99]	Sigma-Delta	12.5~15	0.135~1.92 *	2.6~3.7	65nm
[100]	Sigma-Delta	9~12	20*	3.9~20.1	90nm
[101]	Sigma-Delta	11~14.5	0.2/1/3. 84*	3.4~6.8	90nm

*: signal bandwidth (MHz)

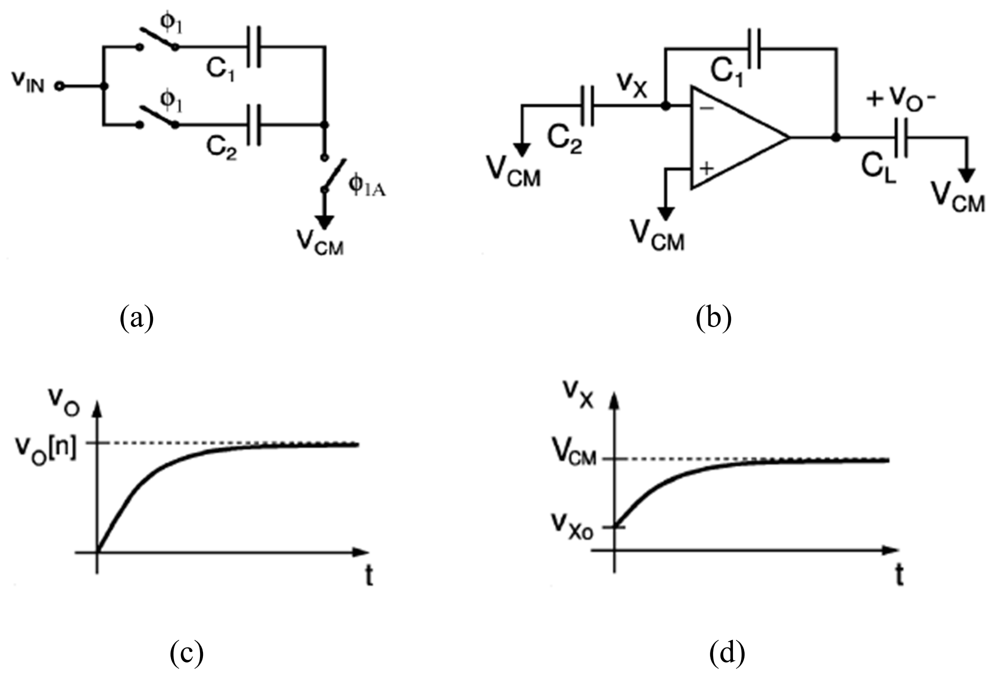
5.2.4 Challenges

Trying to make an ADC “reconfigurable” always results in compromised linearity and/or noise performances, because of the higher order effects induced by extra switches, routing, and control units that are required to realize programmability functions. Therefore, the biggest design challenge is to reduce these degradations, and show comparable power consumption compared with an ADC dedicated to the same performance. An efficient reconfiguration methodology is proposed for medium-high speed ADCs, which will be discussed in detailed in Chapters VI and VII.

5.3 Low Power High Efficiency ADCs

5.3.1 Motivations

Typical switched-capacitor (SC) ADCs requires high gain/high speed Opamps in feedback configuration for precise charge transfer, as shown in Fig. 5.4. The role of the Opamp is to force the virtual ground condition, for the entire charge transfer phase, by driving the output voltage Vo until the virtual ground node Vx equals the common mode voltage Vcm. The accuracy of the output voltage is determined by how well the virtual ground condition is satisfied: a signal-independent error in the virtual ground only generates an offset, which can be eliminated by any auto-zeroing techniques; while a signal-dependent error in the virtual ground results in gain errors and/or nonlinearities, degrading the ADC performance. The accuracy of the virtual ground condition is inversely proportional to the Opamp open-loop gain, therefore the gain must be large enough to ensure a small enough signal-dependent error for a specific application. Finite Opamp gain and insufficient settling are the two main factors that cause signal-dependent errors in the virtual ground condition, thus high gain, high speed Opamps are required.



(a) (b)

(c) (d)

Fig. 5.4. Typical Opamp-based SC gain stage (a) sampling phase (b) charge transfer phase (c) transient response of the output (d) transient response of the virtual ground.

Achieving ultra low power is equivalent to maximizing the ADC power efficiency. The fundamental power limit of an ADC is set by the random, unavoidable electronic noise. However, the precision Opamps introduce more than two orders of magnitude in power overhead [103]. Past attempts to reduce Opamp-based ADC power (e.g. optimum stage scaling, Opamp sharing, etc.) has limited improvement. Ultimately there exists a bound for power dissipation due to the inherently inefficient operation of Opamp-based class-A residue amplifier in the conventional ADC architectures.

Furthermore, Opamp design becomes the most difficult aspect for switched-capacitor circuits in the scaled CMOS process because: 1) shrinking supply voltages results in smaller signal swing, hence larger capacitors to reduce thermal noise and maintain the same dynamic range, thus power increases for the same speed; 2) lower intrinsic gain results in smaller Opamp gain, hence less accuracy. To approach the ultimate power efficiency, we need to get rid of the Opamps!

5.3.2 Techniques to Eliminate Precision Opamps

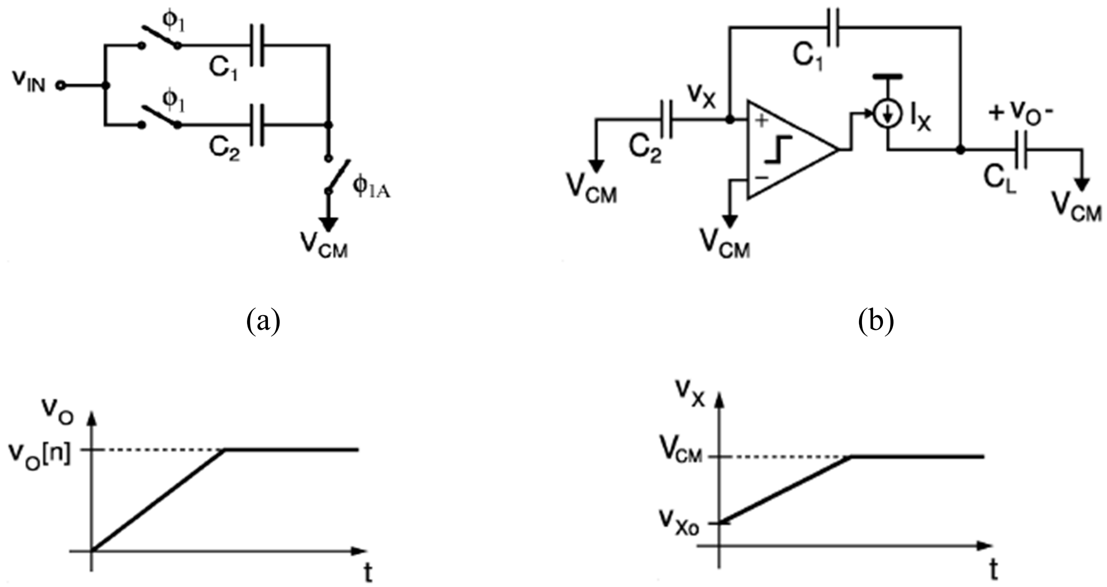
A new family of low power pipelined and sigma-delta ADCs based on inverters [104][105], comparators [106][107], dynamic source followers[108], and charge pumps[109], are recently reported. We will discuss the concepts, the pros and cons for each technique in the next four sub-sections. In addition, advanced digital calibration techniques enable more nonlinear and more imprecise analog parts to be used, such as low gain Opamps in open loop operation with incomplete settling [103],[110].

5.3.2.1 Comparator-based [106][107]

Fig. 5.5 shows a comparator-based SC gain stage. Its sampling phase is the same as an Opamp-based stage; in the charge transfer phase, first, a narrow pulse shorts Vo to ground, and preset Vx to be below Vcm; next, the current source Ix charges up the capacitor network $C_L/C_1/C_2$ to generate a constant ramp on Vo and Vx; the voltage continues to ramp up until the comparator detects the virtual ground condition(i.e. Vx = Vcm) and turn off the current source.

Comparing Fig. 5.5(c) and (d) with Fig. 5.4(c) and (d), we can observe that in the comparator-based gain stage, the voltage settles to the final value in a constant-slope ramp, while in the Opamp-based case, it settles exponentially. In switched-capacitor circuits, the shapes of transient response do not matter, in fact, even two different

Opamp-based systems may have dramatically different transient responses, due to different Opamp performance such as slewing. It is the accuracy of virtual ground condition at the instant when sampling switches turns off determines the charge transfer precision. In an Opamp-based topology, the Opamp forces the virtual ground condition by negative feedback; while in the comparator-based case, the comparator sweeps the output voltage and searches for the virtual ground condition in an open loop scheme; this way, the comparator determines the sampling instant and all charge on C_2 is transferred to C_1; the same charge transfer is realized as in the Opamp-based implementation.



(a) (b)

(c) (d)

Fig. 5.5. Comparator-based SC gain stage (a) sampling phase (b) charge transfer phase (c) transient response of the output (d) transient response of the virtual ground node.

The comparator-based topology has the following advantages:

1) Stability issues due to feedback are removed, because the virtual ground condition is detected in an open-loop manner.

2) The noise bandwidth is about 3-5 times smaller compared to Opamp-based.

3) Input-referred noise power-spectrum-density(PSD) is about 2-4 times smaller than the Opamp-based design.

4) It is easier to generate a constant ramp than designing a high gain Opamp.

Limitations of the comparator-based technique include the following:

1) The finite output impedance of the current source becomes a bottleneck in the design, it causes ramp rate variations hence nonlinear overshoot voltages, which is similar to the finite Opamp gain effect in an Opamp-based ADC; this will results in static integral nonlinearities(INL).

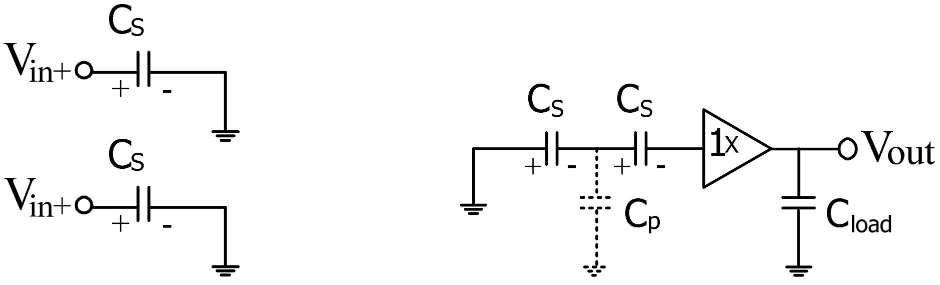
2) Finite comparator delay.

3) Voltage drop across switches due to finite on-resistance.

Digital calibration is needed to compensate for the nonlinearities in order to achieve higher resolution.

5.3.2.2. Charge Pump-based [109]

As shown in Fig. 5.6, this approach is inspired by capacitive charge pumps where successively larger voltages can be obtained by sampling voltages on different capacitors, and subsequently connecting each capacitor in series to yield a total voltage that is the sum of the voltages sampled on each capacitor. Based on this concept, a stage gain of 2X is achieved by passive charge sharing in the open-loop manner. In the sampling phase, Vin is sampled on two capacitors Cs; in the amplifying phase, the two Cs is connected in series to yield an Vout as twice that of Vin. A unity gain buffer is added to prevent charge sharing between capacitors in different stages.



(a) (b)

Fig. 5.6. Charge pump-based SC gain stage (a) sampling phase (b) amplification phase.

Charge pump-based approach has the following advantages:

1) The only active block in the ADC is the unity gain buffer, and a source follower with simple digital calibration is sufficient for 10bit resolution [109]. For higher performance, an Opamp in unity gain feedback can be used; in that case, since the feedback factor is twice that of a traditional Opamp-based stage, same speed can be achieved as traditional approach with half the power consumption.
2) Since the buffer comes after the passive gain block, the buffer noise, when referred to the input, is divided by the amount of passive gain squared; while in a traditional Opamp-based stage, the Opamp noise is not divided by the

stage gain when referring to the input; therefore, the noise from active circuitry in a charge pump-based approach contributes less to the overall noise floor, enabling further power reduction.

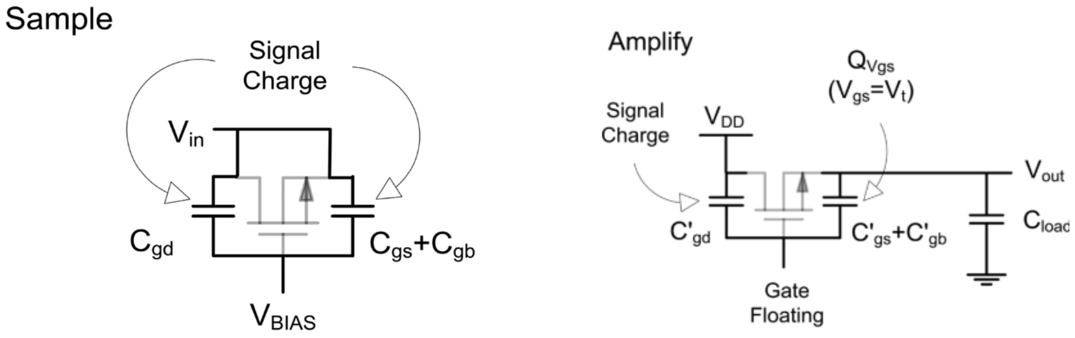
Limitations of the charge pump-based technique include the following:

1) The unity gain buffer becomes the bottle neck for higher speed and higher resolution applications. Larger power is required for the buffer, which limits the power efficiency of the entire ADC. Output swing causes gain variation, which degrades the buffer linearity.
2) Digital calibration is required.

5.3.2.3 Dynamic Source Follower-based [108]

This approach is much more aggressive compared to the previously discussed comparator/charge pump-based methods, because all active circuitries have been eliminated from the signal path.

Fig. 5.7 illustrates the conceptual idea: in the sample phase, the MOSFET is biased in depletion region with the gate tied to V_{BIAS} and the source/bulk/drain tied to Vin; any changes in the incremental input signal will cause incremental changes in the total charge at the MOSFET gate; in the amplification phase, the gate is left floating, the drain is tied to V_{DD}, and the source/bulk is tied to Vout. Consequently, any incremental changes in the total gate charge can only be due to the charge on C_{gd}. The transistor acts as a source follower and it charges the load capacitance C_{load} until it settles, when the drain current approaches zero and $V_{gs} \approx V_t$.



(a) (b)

Fig. 5.7. Dynamic source follower-based SC gain stage (a) sampling phase (b) amplification phase.

This single-transistor residue amplification mimics the charge-redistribution around an Opamp; it dynamically charges Cload without a large bias current. Majority of the supply current is delivered directly to the load, yielding significant improvement in power efficiency.

The advantage of this dynamic source follower-based approach includes:

1) No active circuit in the signal path, thus no dependence on high intrinsic gain amplifiers, making it suitable for future technology scaling.

2) Low input capacitance relaxes the driving circuit power consumption. This is its selling point compared with SAR ADCs.

There is a bottleneck in this technique: since the amplification is based on the charge distribution of a simple transistor, the MOS capacitance nonlinearity and modeling inaccuracy limit the achievable resolution, i.e. 8-9bits in 90nm CMOS process [108]. Digital calibration is necessary to improve the resolution.

5.3.2.4 Inverter-based [104][105]

Fig. 5.8 shows the DC gain and gain bandwidth (GB) vs. supply voltage for a CMOS inverter. Maximum DC-gain is obtained in the weak inversion region, while the GB increases with supply voltage and saturates in the strong inversion region. Therefore, the inverter should operate at the boundary between the weak and strong inversion regions for simultaneously achieving high DC-gain and wide GB, which can be realized by making $V_{DD} \approx V_{TP}+V_{TN}$, where V_{TP} and V_{TN} are the threshold voltage of PMOS and NMOS respectively.

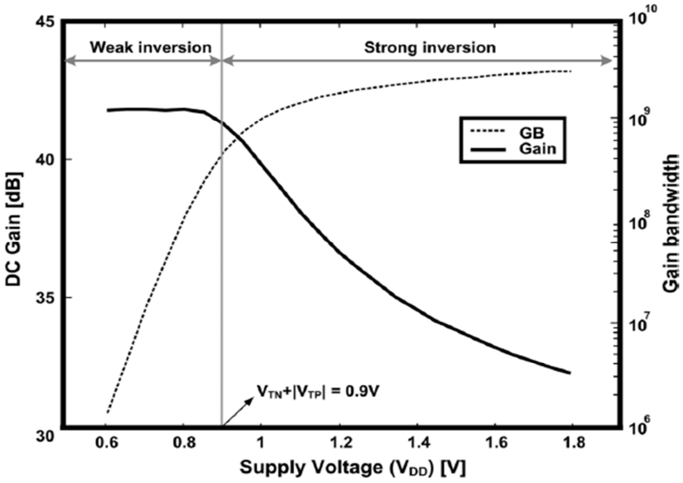
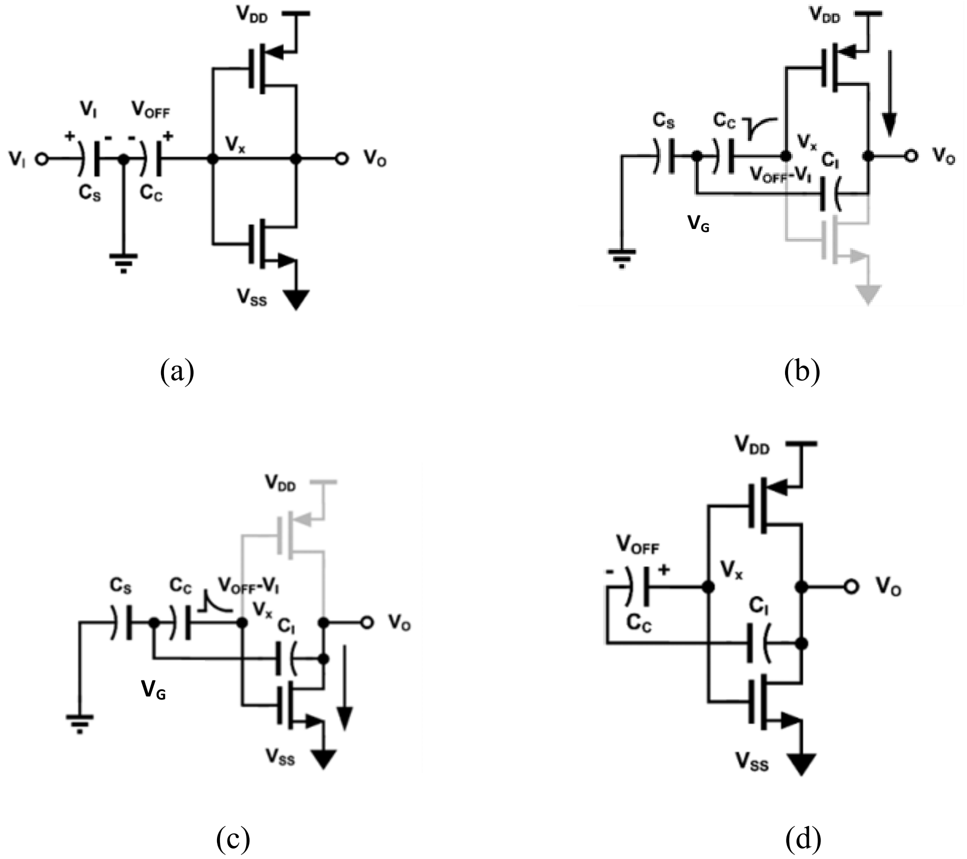


Fig. 5.8. Inverter characteristics: DC gain and gain bandwidth vs. supply voltage.



(a)

(b)

(c)

(d)

Fig. 5.9. Inverter-based switched-capacitor gain stage (a) sampling phase (b) at the beginning of amplifying phase with $V_I > 0$ (c) at the beginning of amplifying phase with $V_I < 0$ (d) Steady state of amplifying phase.

Inverter does not provide inherent virtual ground because it has only one input terminal; when a closed loop is formed, the input node of the inverter is kept close to its offset voltage, thus auto-zeroing technique can be applied to cancel the offset and implement the virtual ground. Fig. 5.9 shows the inverter-based auto-zeroing SC integrator with $V_{DD} \approx V_{TP}+V_{TN}$. In the sampling phase, the inverter is in a unity-gain feedback loop with both transistors operating in the weak inversion region. The offset voltage V_{OFF} and the input V_I is sampled onto Cc and Cs respectively. During the amplifying phase, a negative feedback loop is formed through C_I. Firstly, Vx is instantaneously charged to $V_{OFF} - V_I$ and one of the transistors is biased at strong inversion region, providing high slew rate, while the other transistor is off. Then, Vx gradually returns to V_{OFF}. Because Cc holds V_{OFF}, V_G can be considered as the virtual ground, and the charge on Cs is transferred to C_I. Once the charge transfer is completed, both transistors operate in the weak inversion region again, providing high DC gain with minimum static current.

The inverter-based technique has the following advantages:

1) Compatible with very low supply voltage design.

2) Much lower noise level; the thermal noise is about one-fifth that of a conventional OTA-based topology.

3) Maximized power efficiency for low frequency applications.

Its limitations include the following:

1) Highly sensitive to supply voltage variations (e.g. 10% V_{DD} variations would result in 85% variations in current consumption.

2) Sensitive to threshold voltage variations(e.g. $\pm$ 5% V_{th} variations results in 13% variations in GB)

3) Inverter has lower gain and GB compared to an Opamp, thus generally requires digital assisted when applying it to pipeline ADCs. But it is suitable for sigma-delta ADCs due to the relaxed specifications for analog blocks.

Table 5.2 compares the above discussed four techniques. The dynamic source follower-based technique is the best when considering both the power efficiency and compatibility with technology scaling.

5.3.3 Renaissance of SAR ADCs

ADCs can be categorized into two big clusters: 1) Opamp-limited ADCs, including pipeline ADCs and sigma-delta ADCs; 2) comparator-limited ADCs, including SAR ADCs and flash ADCs. For the first type, the power-hungry precision Opamps introduce more than two orders of magnitudes in power overhead, thus a few techniques have been reported to substitute the Opamp, as have been discussed in section 5.3.2. A recent renaissance of SAR ADCs [111]-[114] confirms the trend of eliminating Opamps and the advantages of comparator-limited ADCs. The only active block in a SAR ADC is a comparator, which operates as logic gates thus benefits from technology scaling and can potentially achieves ultra-low power consumption. The power dissipation scales with sampling rate, due to the intrinsic characteristic of comparators; the power also scales with resolution and the resolution can be easily reconfigured by controlling the binary search algorithm. These features provide great flexibility.

Table 5.2. Comparison of opamp-less ADCs

Technique	Active Elements in Signal Path	Bottleneck	Power Efficiency	Technology Scaling?
Comparator - based	Comparator + Current source	Comparator delay; current source output impedance	🙂	😐
Charge pump-based	1x buffer	Buffer noise; buffer nonlinearity	😐	🙁
Dynamic source follower-based	Single MOS transistor	MOS cap nonlinearity; higher order effects	🙂	😀
Inverter-based	Inverter	Sensitive to PVT variations	🙂	🙂

Table 5.3 summarizes the performances of state-of-art high efficiency ADCs. Inverter-based sigma-delta [104][105] can achieve high resolution, other techniques [106]-[109]requires digital calibration circuitry for higher resolution. But as the digital circuit power goes down with scaling technology, higher power efficiency will be obtained. SAR ADCs are promising in terms of power efficiency and amenability to technology scaling.

Table 5.3. State-of-art low power high efficiency opamp-less ADCs

Ref.	Architecture	Resolution	Sampling rate	Power	Efficiency	Process
[104]	Inverter-based sigma-delta	13bit	50-300MHz	950µW	0.3pJ/ step	65nm
[105]	Inverter-based sigma-delta	14.2bit	4MHz	36 µW	0.098pJ/ step	0.18µm
[106]	Comparator-based Pipeline	8.6bit	7.9MHz	2.5mW	0.8pJ/step	0.18µm
[107]	Comparator-based Pipeline	10bit	50MHz	4.5mW	0.088fJ/step	90nm
[108]	Source follower-based pipeline	9.4bit	50MHz	1.44mW	0.119pJ/step	0.13µm
[109]	Charge pump based Pipeline	10bit	50MHz	9.9mW	0.3pJ/step	0.18µm
[111]	SAR	12bit	100kHz	25µW	0.165pJ/step	0.18µm
[111]	SAR	9bit	40MHz	820µW	0.054pJ/step	90nm
[112]	SAR	10bit	1MHz	1.9µW	0.0044pJ/step	65nm
[114]	SAR	9.4bit	100kHz	3.8µW	0.056pJ/step	0.18µm

5.4 Time Domain ADCs

Another trend in ADC design that leverages the strength of ultra-deep-submicron technologies is time-domain ADCs. The main concept is to represent information by the time difference between two edges, i.e. pulse width, instead of by voltage difference (see Fig. 5.10), and the minimum detectable time-step correspond to an LSB. The quantization is performed in the time domain instead of voltage domain, by a time-to-digital converter (TDC).

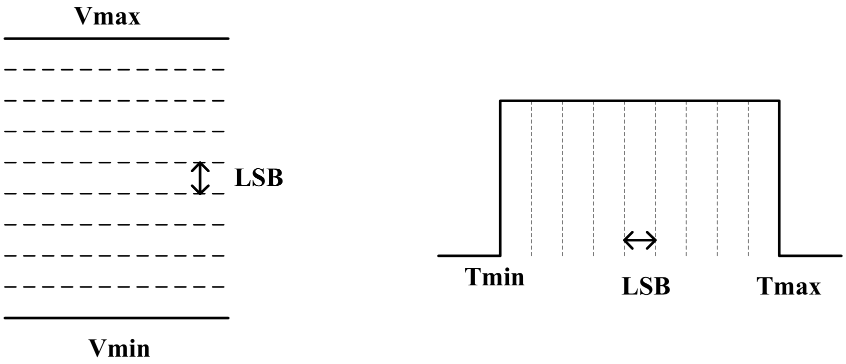


Fig. 5.10. Time-domain vs. voltage-domain operation.

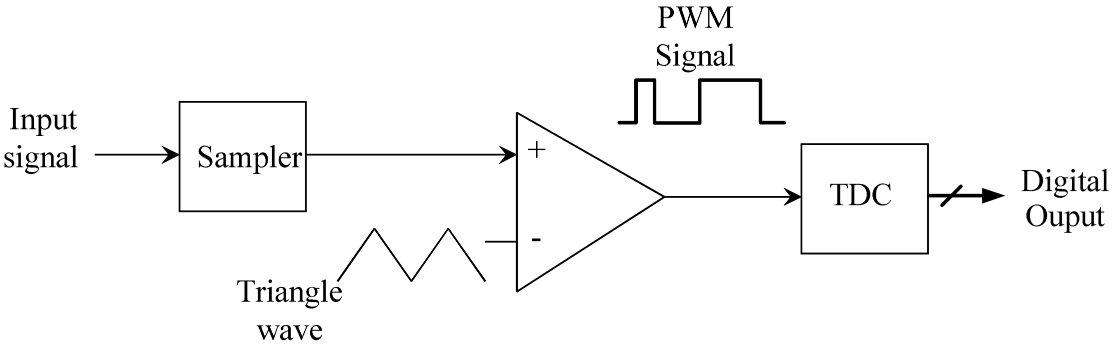


Fig. 5.11. Block diagram of a time-domain ADC.

Fig. 5.11 shows the block diagram of a time-domain ADC. The input signal is sampled using a simple sample-and-hold circuit. A pulse-width-modulated (PWM) generator converts the voltage signal into a time-domain signal, and then a TDC performs quantization and generates the digital outputs.

Why do we need time-domain ADCs? As technology scales down, more mismatches in smaller-size devices, shrinking supply voltage, increased device noise and nonlinearity, all deteriorate voltage resolution hence reducing the dynamic range, making it harder to design a high resolution ADC.

On the other hand, the dynamic range of a time-domain ADC is defined by the ratio between the largest pulse width that the system can afford and the minimum time-resolution provided by a specific process. For conventional TDCs the minimum time-resolution is one-inverter-delay, which is less than 10ps in the state-of-art process, and the time-resolution keeps improving as process scales. The maximum pulse width is specified by two factors: 1) input signal bandwidth. For example, for an audio signal with 25 KHz bandwidth, the minimum sampling rate is 50 KHz, corresponding to a maximum pulse width of 20μsec when converted to a PWM signal; hence the dynamic range can be as high as 126dB. 2) Accumulated jitter introduced by the TDC also limits the maximum pulse width. In order to have a resolution of one-inverter delay when quantizing the pulse width, the standard deviation of the accumulated jitter σ_j, through the duration of the pulse, should be kept below the time-resolution, i.e. smaller than one inverter delay.

One advantage for time-domain ADCs is that the SNR is limited by the timing jitter instead of the voltage noise level of the quantizing device. Timing jitter is defined as the voltage noise divided by the slew rate (SR) of the edge transition. Although the noise increases with technology scaling, the SR increases much faster, resulting in a better jitter performance of the TDCs. Consequently, same performance can be obtained with smaller power consumption as process scales.

A few techniques have been developed to achieve timing resolution of sub-inverter-delay in order to get a larger dynamic range. “Vernier delay line”[118], one of the most well-known techniques, obtains a minimum detectable resolution of the difference between the delays of two differently-sized inverters. The main drawback of this technique is that the signal latency increases as the resolution improves, which makes it only suitable for single shot conversion or small bandwidth signals.

In [119], a local-passive-interpolation (LPI)-based TDC was reported that overcame the latency problem. It interpolates the rising edges of the inverters in order to generate intermediate edges using passive components. The main advantage of this technique is its fast conversion compared to Vernier-TDC; furthermore, the interpolation depends on the ratio between the passive components, making it robust against the global mismatches.

Another technique that alleviates the latency problem is the multistage TDC [120], in which multistage pulse quantization is used. In the first step, a chain of buffers are employed to perform a coarse quantization similar to the conventional TDC. In the second step, a fine quantization is performed using Vernier delay line. This way, it is the resolution of the conventional TDC that determines the maximum input pulse width for the Vernier line, which corresponds to a small and compact Vernier line. Advantages of this technique include: 1) area-efficient. 2) the power consumption per conversion is much smaller than the pure Vernier TDC as the number of stages is much less. 3) Latency as low as less than one buffer delay can be achieved. Limitations of this include: 1) two delay locked loops, instead of one, are required to calibrate the delay of each stage. 2) the MUX used to multiplex the first level signal may induce dead zone in the signal, leading to a degradation in the overall resolution. A resolution of 10ps was reported using this technique in 0.18μm technology.

Other techniques for obtaining a sub-gate-delay timing resolution include the sub-gate-delay-based TDC [121], in which many inverters with different sizes are in parallel such that they can have different delays. As scaling the inverter's size induces an additional delay that is usually less than the single-gate delay, higher resolution can be obtained. A comparable technique is the time-shrinking-delay-line TDC [122], which uses a single delay line to digitize the signal. In order to increase the timing resolution, the delay elements are designed such that the pulse shrinks while propagating through the line. The pulse is also used to trigger the flip-flops connected to the delay element outputs and changes their state. As the pulse propagates through the line, the pulse width decreases until it vanishes. When the pulse vanishes, the remaining flip-flops will not be triggered and its old state will be maintained indicating that the pulse vanished. The attainable resolution depends on the pulse-width-shrinking. This technique suffers from the large latency as the Vernier delay line-based TDC.

Table 5.4 compares different TDC architectures and shows the tradeoff between resolution, area, power, design robustness and latency. It is clear that designing a TDC with high resolution, low latency, and good robustness is hard; the performance limitation from TDC together with PWM nonlinearity limits the achievable resolution for an open-loop time-domain ADC. In order to break this trade-off, a close-loop TDC-based ADC is proposed [116], as shown in Fig. 5.12. The main idea is to incorporate a conventional TDC that acts like a multi-bit quantizer in conventional voltage mode ADCs. A simple TDC is employed, with a 80ps resolution corresponding to 4.5bits. The quantization error from the TDC is shaped by the negative feedback loop as in any sigma-delta ADCs; the loop also shapes the nonlinearity from PWM, ending up in 10bit resolution for the entire ADC with 20MHz bandwidth. The bottleneck of the design is to provide a feedback pulse with sub-ps accuracy from the TDC, which was proven feasible in 65nm technology [116].

Table 5.4. Comparison of the different TDC technique

	TDC Architecture	Resolution	Latency	Area	Robustness	Power
[117]	Conventional	low	low	compact	moderate	small
[118]	Vernier	high	large	large	poor	large
[119]	Local passive interpolation	moderate	low	medium	good	small
[120]	Multistage	high	low	medium	moderate	moderate
[121]	Sub-gate delay	high	low	large	poor	large
[122]	Time shrinking	low	large	large	poor	large

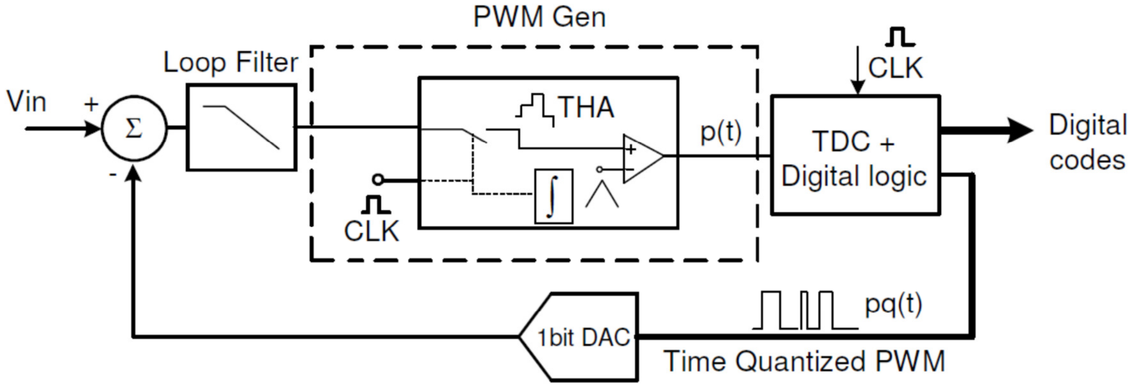


Fig. 5.12. Block diagram of the closed-loop TDC-based ADC.

5.5 Proposed Minimum Current/Area Implementation of Cyclic ADCs

5.5.1 Introduction

Cyclic ADC and successive approximation (SAR) ADC are two popular ADC types for medium resolution, medium speed applications. They both belong to “Bit-at-a-time” ADCs, i.e. multi-step ADCs that resolves one bit per step, and requires multiple conversion steps to generate one digital word. Fig. 5.13 shows the block diagram of a cyclic ADC, it is essentially the same as pipeline, but a single stage is used in a cyclic fashion for all operations. Fig. 5.14 illustrates the operation of SAR ADCs. Its internal digital-to-analog converter (DAC) is initially set to midscale for the comparator to resolve bit 1(MSB), and the output is stored in the SAR logic, which controls the DAC to set to ¼ or ¾ for the second comparison step. This binary search continues until the LSB is resolved. For 10-bit resolution, a capacitor DAC takes more area than a resistor DAC. The SAR ADC with an R-2R DAC requires 30 resistors of 10-bit matching accuracy, these together with binary scaled CMOS switches consumes large area; on the other hand, the conventional cyclic ADC with a multiply-by-two gain stage only requires 8 capacitors and two OTAs in the signal path. Furthermore, capacitors inherently have better matching than resistors, making the cyclic ADC more area efficient than SAR ADC at a 10-bit level. In this section, techniques are proposed to further reduce the area and power consumption of a cyclic ADC.

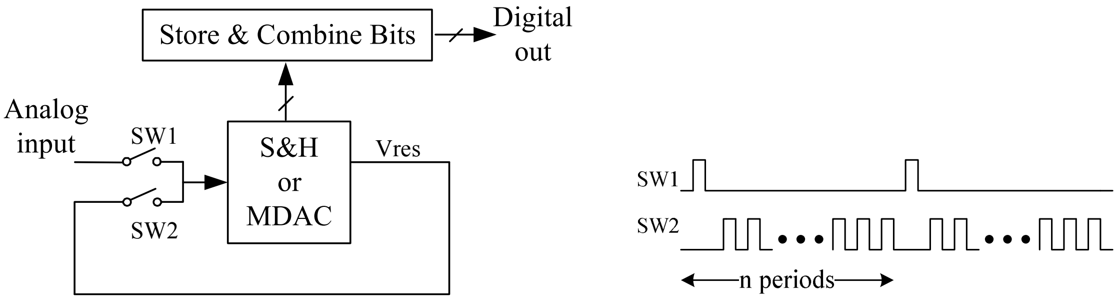


Fig. 5.13. Block diagram of a cyclic ADC.

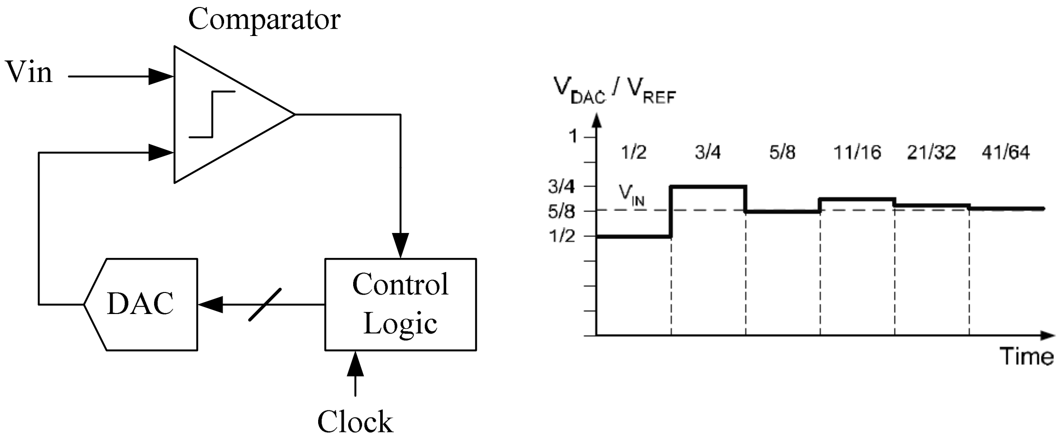


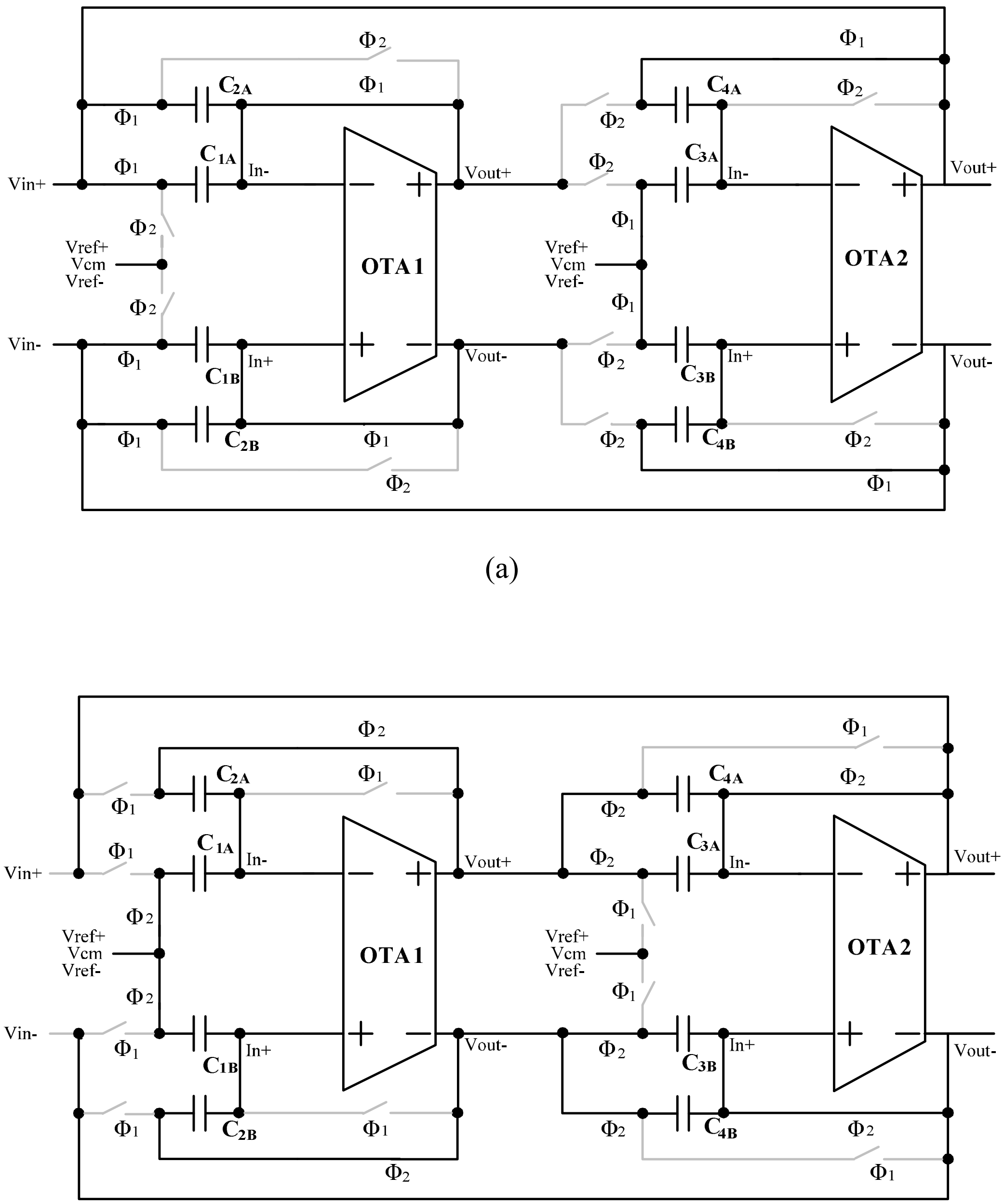
Fig. 5.14. Block diagram of a SAR ADC.

5.5.2 Proposed Solution

In a cyclic ADC, the residue signal is cyclic, thus only one gain stage is needed, and 10 periods are required to convert a 10-bit digital code. In this work, the Redundant Sign Digit (RSD) technique [123] is adopted to enable the use of a cheap comparator. OTA sharing technique [124] is employed to further cut down the power and area.

Fig. 5.15 shows a conventional multiply-by-two gain stage. OTA1 and OTA2 work in an interleaving manner to produce one bit per conversion cycle; in Φ1, capacitors C_{1A}, C_{1B}, C_{2A} and C_{2B} sample the input, OTA1 works in unity gain feedback configuration for offset cancellation, and OTA2 is in capacitive feedback configuration to transfer charge from C_{3A} and C_{3B} to C_{4A} and C_{4B} respectively; in Φ2, OTA1 and OTA2 exchange the role of operations.

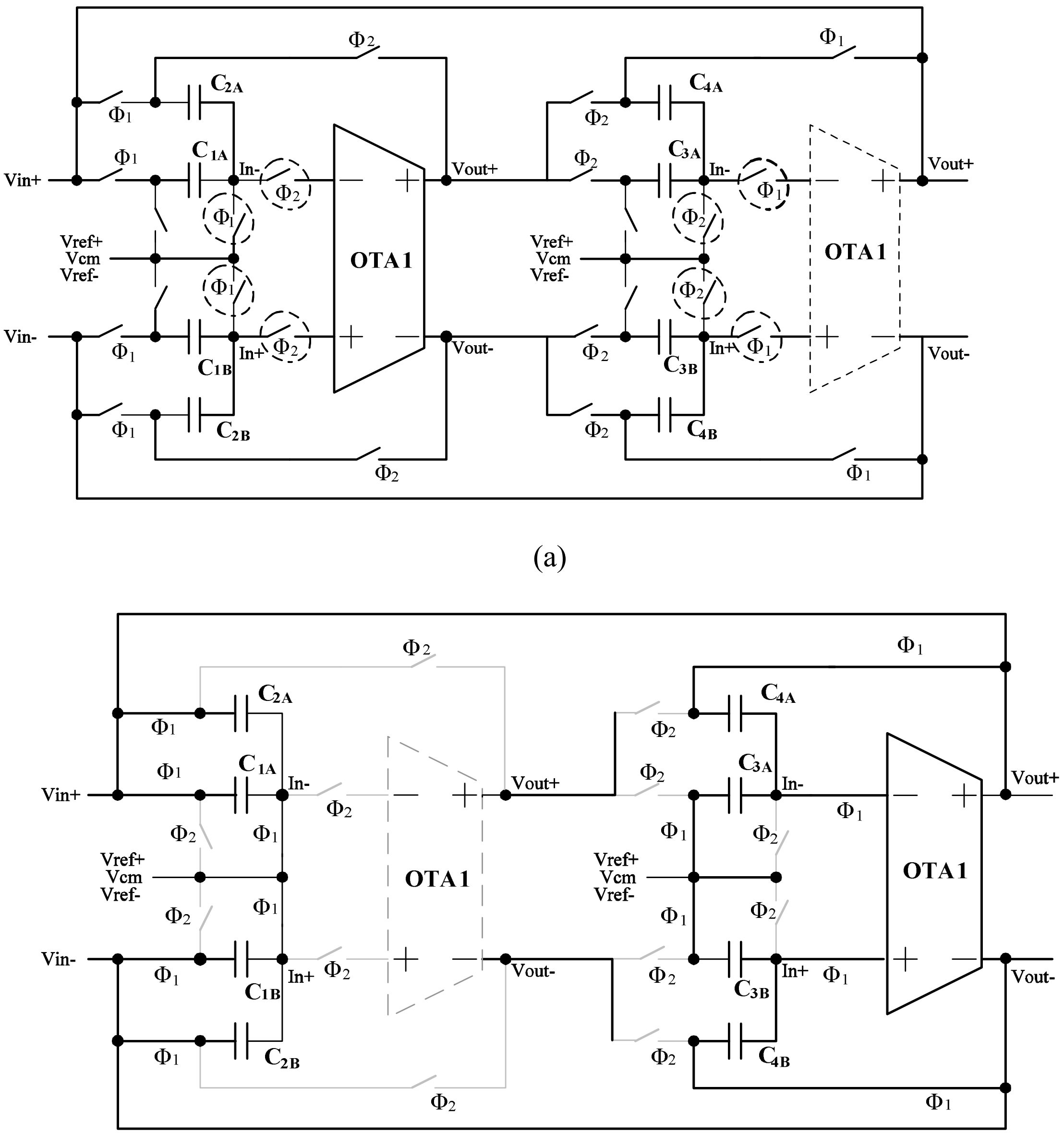
Fig. 5.16 shows a multiply-by-two gain stage with OTA sharing. Eight additional switches (circled) are added compared to the conventional case. Only OTA1 is needed, thus power is theoretically cut by half by eliminating OTA2. However, since OTA1 is always in capacitive feedback configuration for both phases, there is no time for offset cancellation, and the negative terminal of the sampling cap has to be connected to Vcm during sampling phase, instead of virtual ground as in the conventional case. Therefore, offset and flicker noise cannot be stored and cancelled. Offset and flicker noise will not affect current comparator output because of the RSD technique, but they will propagate to the following conversion periods, potentially affecting the final conversion result. Offset translates into a fixed amount of up/down shift in the digital output signal, and is tolerable in certain applications; but flicker noise causes a varying shift, therefore must be cancelled, especially for servo applications where the signal BW is close to DC.



(a)

(b)

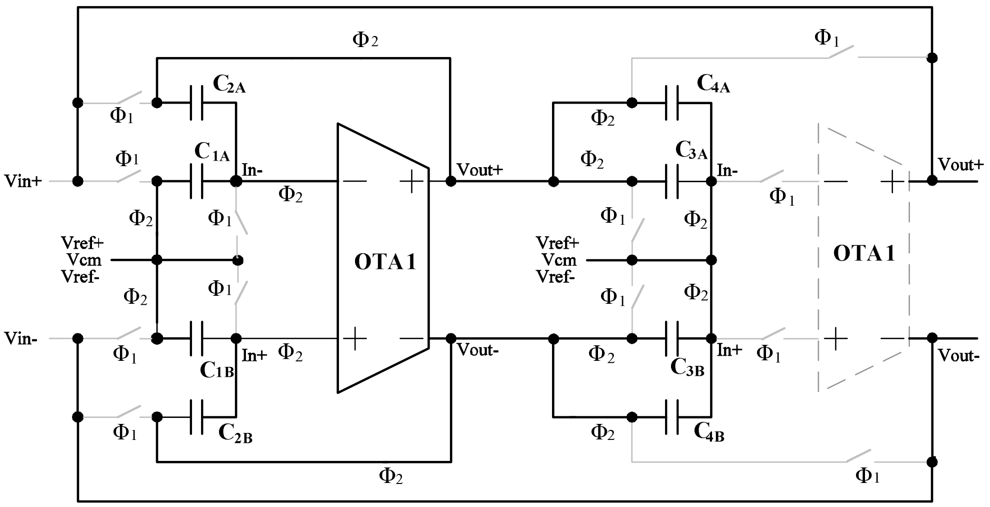
Fig. 5.15. Conventional multiply-by-two gain stage (a) OTA1 in sample phase, OTA2 in charge transfer phase(b) OTA1 in charge transfer phase, OTA2 in sample phase.



(a)

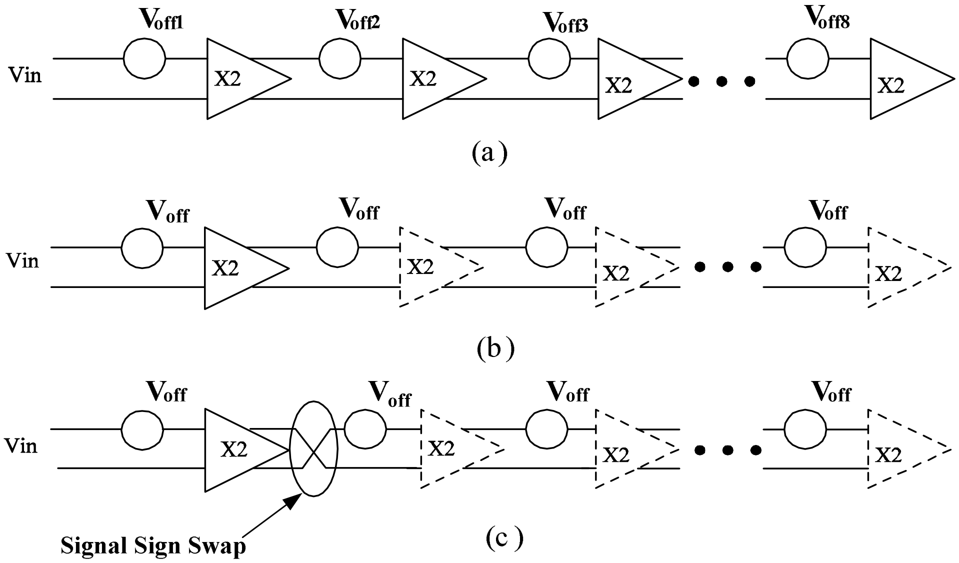
(b)

Fig. 5.16. Multiply-by-two gain stage with OTA sharing technique (a) circuit implementation (b) Φ_1 switches "ON", Φ_2 switches "OFF" (c) Φ_1 switches "OFF", Φ_2 switches "ON".



(c)

Figure 5.16 Continued



(a)

(b)

(c)

Fig. 5.17. Proposed global offset cancellation scheme (a) Pipeline ADC (b) Cyclic ADC (c) Cyclic ADC with global offset cancellation.

Fig. 5.17 illustrates the proposed global offset/flicker noise cancellation technique. For comparison purpose, a conventional pipeline, a conventional cyclic, and a cyclic with proposed global offset cancellation technique are shown. Each triangle (X2) represents a multiply-by-two gain stage. In a conventional 10-bit 1.5-bit/stage pipeline ADC with a two-bit flash last stage, the offset voltages from eight different OTAs are uncorrelated, denoted as V_{off1}, V_{off2}, …, V_{off8}, the total input referred offset is:

$$V_{Total\_OS\_pipeline}=\sqrt{V_{off1}^{2}+\left(\frac{V_{off2}}{2}\right)^{2}+\left(\frac{V_{off3}}{2^{2}}\right)^{2}+\cdots+\left(\frac{V_{off8}}{2^{7}}\right)^{2}} \tag{5.1}$$

Fig. 5.17 (b) shows a cyclic ADC with OTA sharing, since a single OTA is reused for all the conversion cycles, the input referred offset in each cycle is correlated, denoted as V_{off}. The total input referred offset is:

$$V_{Total\_OS\_cyclic}=V_{off}+\frac{V_{off}}{2}+\frac{V_{off}}{4}+\cdots+\frac{V_{off}}{256}\cong 2V_{off} \tag{5.2}$$

Based on this observation, we proposed the global offset cancellation technique, as illustrated in Fig. 5.17 (c). By doing just a signal sign swap after the first multiply-by-two operation, the total input referred offset of the ADC is reduced by 512 times as:

$$V_{Total\_OS\_cylic}=V_{off}-\frac{V_{off}}{2}-\frac{V_{off}}{4}-\cdots-\frac{V_{off}}{256}=\frac{V_{off}}{256} \tag{5.3}$$

5.5.3 Simulation Results

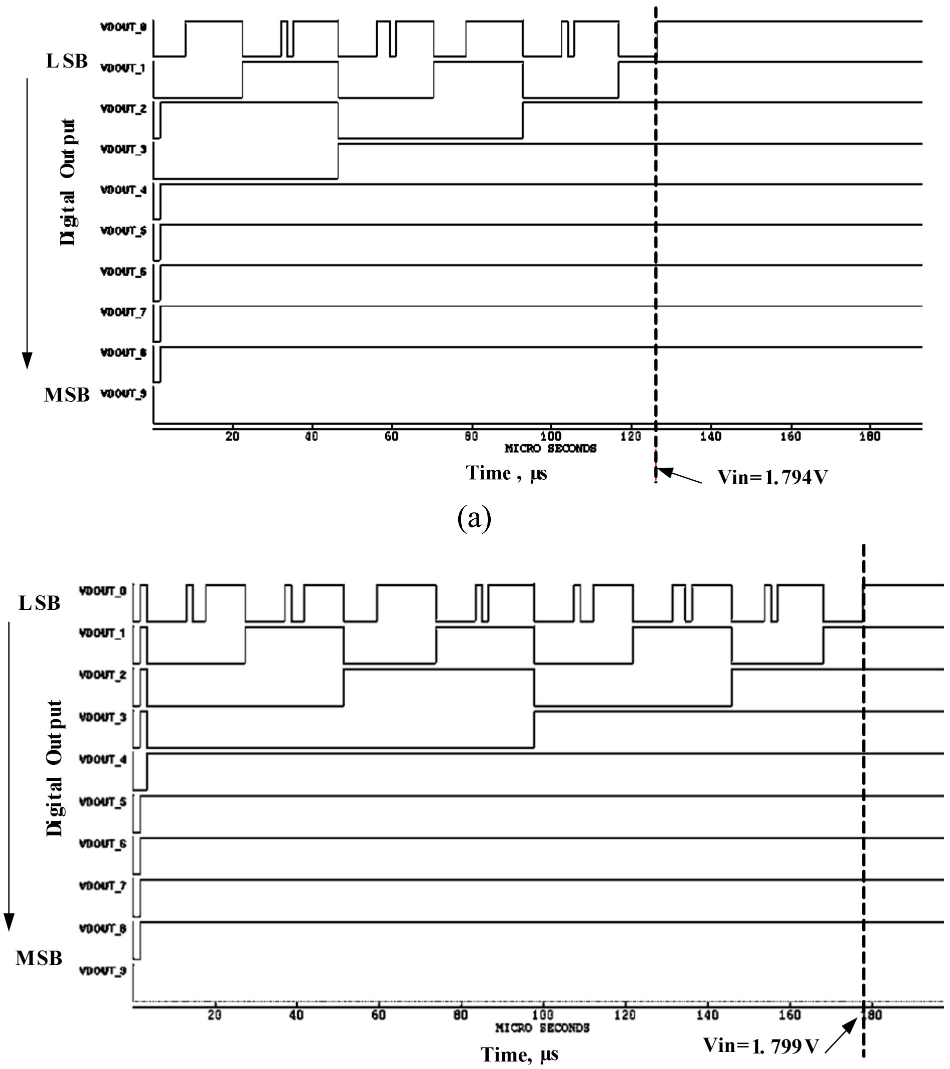
The whole cyclic ADC is designed at transistor level in TI 0.35μm CMOS process, incorporating the RSD technique, the OTA sharing technique, and the proposed global offset cancellation technique. The 10-bit 1MS/s ADC occupies an area of 0.14mm^2, and consumes 2.3mW from a 3.3V supply. An NMOS input folded-cascode OTA with gain boosting and switched-capacitor common-mode feedback (CMFB) is used. A 5mV input referred offset is extracted from intensive Monte Carlo mismatch simulations for the OTA and intentionally added in the simulation to verify the proposed offset/flicker noise cancellation technique. The 10-bit ADC input range is 0.6V~1.8V, thus a 1.17mV input increment corresponds to a change in the digital output code. A slow ramp of 0.6V~1.8V is fed into the ADC. Each digital code is made to ideally appear ten times. Fig. 5.18 shows the 10-bit digital output corresponding to the input section near 1.8V. Fig. 5.18 (a) shows that without offset cancellation, the digital output becomes all "1" when Vin ≈ 1.794V, therefore offset has caused the ADC output range to shift up; after introducing the global offset cancellation technique, as shown in Fig. 5.18 (b), the digital output becomes all "1" when Vin ≈ 1.799V, confirming an accurate ADC output range. Note that although offset does not affect DNL, flicker noise is varying slowly and will degrade DNL, hence must be cancelled. Since flicker noise can be treated as constant DC offset within one conversion period, the above simulation showing offset cancellation also verifies the effectiveness of flicker noise cancellation.

5.5.4 Summary

Global offset cancellation technique is proposed to alleviate the offset and flicker noise problems arising from OTA sharing in a cyclic ADC. This is most beneficial to high volume, cost/area sensitive product lines, such as servo application. The proposed cyclic ADC has been fabricated as part of the servo chip for measuring and controlling the spindle power for silent motor rotation. Although the standalone ADC was not characterized, its functionality has been verified at the system level in silicon. As shown in Table 5.5, the small area and low power consumption of this proposed cyclic ADC results in better FOM (as defined in equation (5.1)) when compared to the state-of-art ADCs[125][126] in the same 0.35μm CMOS process.

5.6 Conclusions

This chapter has projected three new trends for next generation ADCs. The speed/resolution/power reconfigurable ADCs present more intelligent energy utilization, and is suitable for the future adaptive systems. ADCs without power-hungry precision Opamps can potentially approach the fundamental power limit and be applied to ultra-low power applications. The increasing time-resolution associated with technology scaling makes the time-domain-based ADCs very attractive over conventional voltage-domain-based, for the next generation ADCs.



(a)

(b)

Fig. 5.18. 10-bit ADC digital output with a slow ramp input
(a) without global offset cancellation (b) with global offset cancellation.

Table 5.5. Results Comparison with state-of-art ADCs in 0.35µm CMOS process

	Topology	Resolution	Speed	DNL (LSB)	Power (mW)	CMOS Process	Area (mm^2)	FOM
This work	Cyclic	10bit	1MS/s	0.4	2.3	0.35µm	0.14	2.2
[125]	SAR	7bit	0.1MS/s	0.45	0.2	0.35µm	0.15	15.6
[126]	Pipeline	10bit	2MS/s	0.5	39	0.35µm	2.24	19

CHAPTER VI

PROPOSED SPEED RECONFIGURABLE POWER SCALABLE ADC

6.1 Introduction

The emerging multi-format video processors and multi-standard wireless receivers have created a great demand for integrating multiple design specifications into a single chip [127][128]. Flexible RF and analog baseband blocks that can meet various specifications with minimum hardware implementation are required in such systems. An "adaptive figures of merit (AFOM)" is proposed in [127]. When it comes to ADCs, a power- and area-efficient reconfigurable ADC with variable bandwidth and dynamic range is a promising solution [85]-[96], [129]-[131]. Customized ADCs have power optimized for only one specification, while a reconfigurable ADC can scale its power at different specifications, enabling minimal power consumption over a broad range of sampling rates and resulting in a more power-efficient design.

On the other hand, time-to-market pressure and increased design complexity create a "design gap" for SoCs. The "design-reuse methodology" has been successfully applied to digital systems; therefore the analog part of the SoC dominates the overall design time, cost, and risk. The ADC is one of the most important analog units, and a reconfigurable ADC provides IP reuse, which can be targeted for a wide range of applications with different specifications, thus reduces design efforts, development costs, and time to market.

This work targets display and imaging systems. Most multi-format video processors for HDTV, SDTV, and PC graphic require a constant resolution with an ENOB > 9bit for accurate color reproduction. However, various sampling rates and effective resolution bandwidth (ERBW) are specified since they are proportional to the number of pixels and refresh rate which are different among different standards. Therefore, although an ADC can be configured in both the resolution and sampling rate, programming the sampling rate is more important and challenging in our target applications. Power/speed configurability is a desirable feature for ADCs targeting energy-constrained applications. A power scalable architecture allows sampling rate programmability while maintaining almost constant power/speed ratio [130]. Table 6.1 summarizes the ADC requirement for component video, PC graphic, and some popular communication standards. The wide variations in the sampling rate requirements, 1MSPS-200MSPS, makes it very challenging to design a reconfigurable ADC covering all these standards.

The Sigma-Delta ADC is an attractive solution for multi-standard wireless receivers design [97]-[101], [132]. It can be configured to achieve larger bandwidth with lower resolution or smaller bandwidth with higher resolution by programming its digital decimation filter. However, the over-sampling feature limits the use of a Sigma-Delta ADC in wide bandwidth applications, such as video processors, since the high sampling frequency results in a high power consumption.

On the other hand, the pipeline ADC has inherently higher operating speed, thus it is more suitable for medium to high speed applications. Furthermore, its sampling rate

Table 6.1. Summary of the ADC specifications

(a) Communication

Standard	Sampling rate (MSPS)*	Resolution(bit)
GPS	4	10
WCDMA	8	9-10
WLAN	44	8-10
WiMAX	10~40	8-10

(b) Component video

Standard	Pixel rate (MSPS)
480p	8.1
480i	18.41
576p	20.736
576i	20.736
720p	22.12 – 55.3
1080p	49.77 – 124.42
1080i	49.72 – 124.30

(c) PC graphic

Standard	Pixel rate (MSPS)**
VGA	21.5
SVGA	28.8
XGA	47.19
XGA+	59.72
SXGA	78.64
SXGA+	88.20
UXGA	115.20
QXGA	188.74

and resolution can be programmed independently, which is desired for multi-format video processors where various sampling rate are needed while the same resolution is required. However, a pipeline ADC is more difficult to program than the sigma-delta. Therefore, this research work explores an efficient implementation of a reconfigurable

ADC based on the pipeline architecture with medium to high sampling rate to cover all the video standards.

Section 6.2 addresses the challenge of analog power scaling, and discusses reconfiguration methodologies for ADCs. Section 6.3 presents the proposed reconfigurable ADC architecture. Section 6.4 describes the circuit implementation for each building block. Layout considerations and measurement results will be presented in the next chapter.

6.2 ADC Reconfiguration Methodology

6.2.1 Analog Power Scaling

For a power-optimized ADC, just enough power is consumed to ensure the required accuracy at a specific clock frequency. The ADC FOM, as defined in (5.1), is proportional to the power/speed ratio; therefore, it is essential to have good power scalability when programming the speed in order to keep a comparable FOM with dedicated ADCs at each setting.

To explore the power scalability, we can recall the power consumption expressions for digital and analog circuits shown below.

$$Power_{\text{digital}} = \frac{1}{2} C V^2 f_s \tag{6.2}$$

$$Power_{\text{analog}} = V \cdot I \tag{6.3}$$

where V is the supply voltage, C is the load capacitance, and I is the total current drawn from the supply. For digital circuits, the average power automatically scales with sampling frequency. In analog circuitry, the power is not an explicit function of frequency. Furthermore, V is kept constant in most cases. To make the power track the clock frequency, it is desirable to make the current as a function of frequency, i.e.: $Power_{\text{analog}}(f_s) = V \cdot I(f_s)$.

6.2.2 Bias Current Scaling

A straightforward way to make the power track the sampling frequency f_s is to scale the biasing current of the active building blocks (i.e. OTAs) [86], [90]-[93], [129]. Fig. 6.1 shows a simple scalable bias current generator.

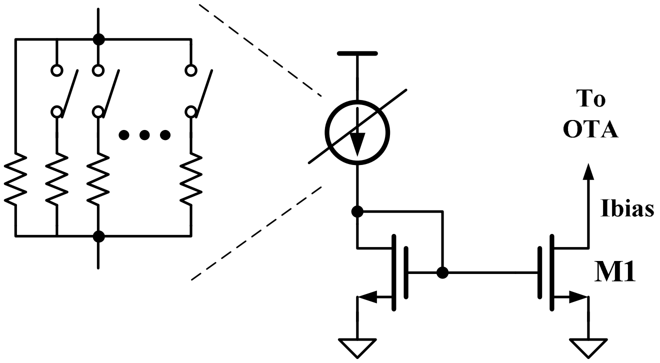


Fig. 6.1. A scalable bias current generator.

For a rough estimation ignoring slewing, f_s is proportional to the closed loop gain bandwidth (GBW) of the OTA in the sample-and-hold (S&H)/Gain stage for a certain settling accuracy. Assuming a single-stage OTA is employed, $GBW \propto g_m / C \propto \sqrt{I}$; here g_m is the transconductance of the input differential pair, C is the total capacitance that the

OTA has to drive (including the capacitive feedback network and the load capacitor). Therefore, if a wide programming range in f_s is required, the current I needs to be scaled by a large ratio. Large variations in bias current will drive the transistor far away from the optimum/intended operation region, resulting in a poor yield.

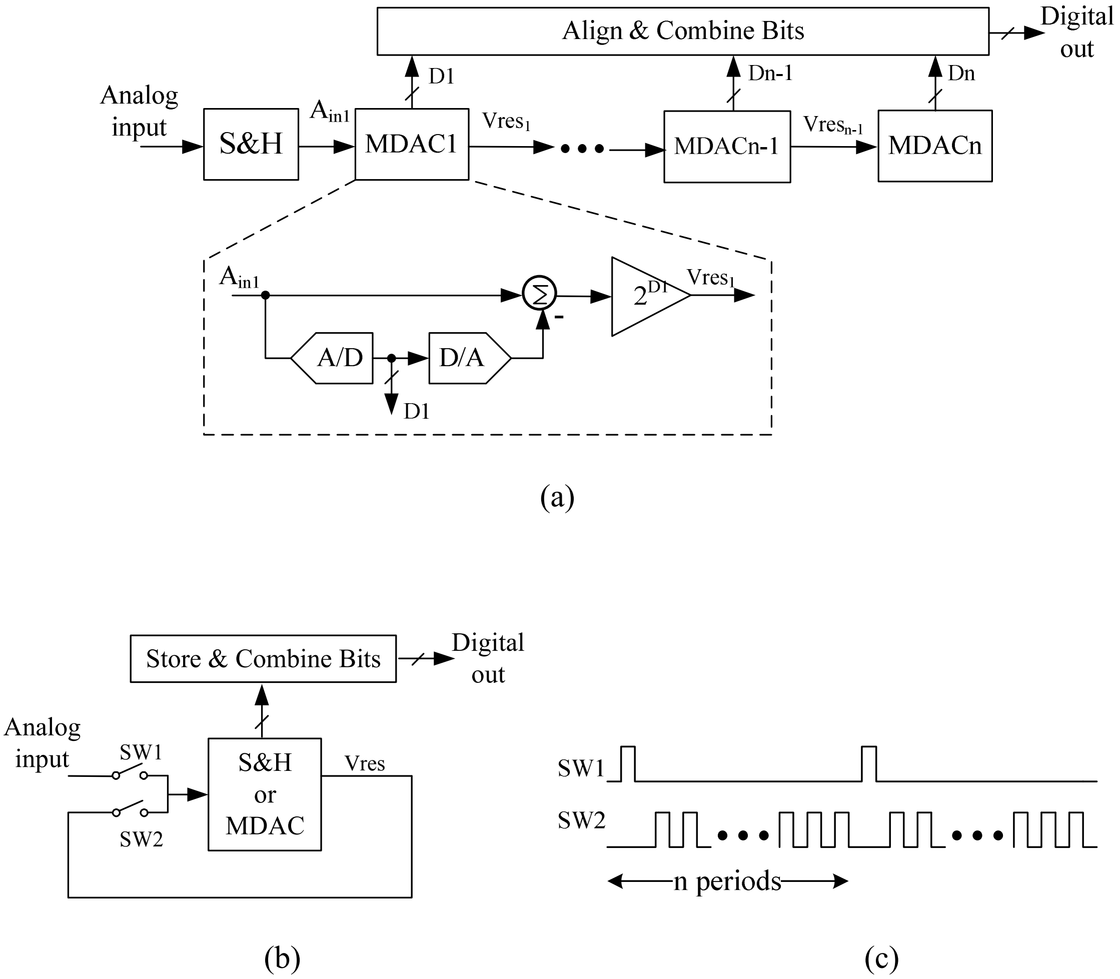
Furthermore, in our particular application, it is desired to configure only the speed while keeping the resolution constant, therefore we need to maintain a constant DC-gain over a large range of bias current. However, bias current variations affect the open-loop DC gain and the maximum output voltage swing of the amplifier: both the open-loop gain and maximum output voltage swing typically decrease with increasing bias current, especially for an OTA with cascode stages. Therefore, biasing current scaling makes the design more difficult.

In [90] and [93], which simply employ the bias current scaling method, good power scalability is reported, but with small speed programming ratio (<7). These results indicate that bias current scaling can only achieve very limited speed/power programmability.

6.2.3 Architecture-level Reconfiguration

Each type of ADC has its bounds on resolution and speed [128], as illustrated in Fig. 5.3. Typically, Sigma-delta ADCs covers the low speed, high resolution applications; flash ADCs occupy the high speed, low resolution applications; SAR and cyclic ADCs are suitable for medium resolution, medium speed applications; pipeline ADCs are good for medium-to-high speed, medium resolution applications, while the fast progress in digital calibration techniques have been improving its speed and resolution, enabling pipeline ADCs to break the resolution limit set by sigma-delta, and the speed limit set by flash. A sound reconfigurable ADC design should combine and take advantage of different ADC architectures that share the same building blocks (i.e. minimize overhead) to cover a wider performance range. A reconfigurable ADC can be conceptually viewed as an ADC with a configurable switch matrix, which adjusts the ADC topology to minimize power consumption at each point in the performance space. Trying to make an ADC "reconfigurable" usually results in compromised linearity and/or noise performances, due to the higher-order effects induced by extra switches and control units for programmability functions. Therefore, a big challenge is to reduce these degradations, and show comparable power consumption at each performance node compared with a dedicated ADC. By taking advantage of the similarity between different ADC architectures, we can minimize the modification/additions to the analog part of the original ADC and reuse the switches as much as possible.

The Cyclic ADC (algorithmic ADC) [102] is the ADC type that shares the most similarities with pipeline ADC. Fig. 6.2 shows the conventional diagram of an n-stage pipeline ADC and an n-cycle cyclic ADC. The Multiplying DAC stage (MDAC) in both the pipeline and cyclic has the same building blocks, i.e. a sub-ADC, a sub-DAC, and a residue amplifier. The difference is: the pipeline ADC passes the residue voltage (Vres) from one stage to the next; while the cyclic ADC recycles the residue back to the input



(a)

(b)

(c)

Fig. 6.2. Block diagram of (a) pipeline ADC (b) cyclic ADC.

of the same stage: in the first conversion period, SW1 is “ON”, and the MDAC stage, configured as S&H, samples the analog input signal; for the next n-1 conversion periods, SW2 is “ON” while SW1 remains “OFF”, and the stage samples its own residue output.

Important observations can be drawn based on the comparison between pipeline and cyclic ADCs: 1) the pipeline and cyclic ADC share the same building blocks; 2) the pipeline ADC is fast since it has n stages working concurrently; the cyclic ADC is n times slower than the pipeline ADC because it has only one stage doing the job of n stages in n sequential cycles with smaller average power per conversion cycle; 3) the cyclic ADC is not power efficient because the hardware needs to be designed for MSB accuracy with respect to noise, settling, and linearity, while "stage scaling" can be applied for a pipeline ADC to relax the requirements of the stages along the chain. Stage scaling is a typical technique for power-efficient pipeline ADC design, by scaling down the biasing current of the active blocks (i.e. OTAs) and the capacitors at a proper ratio, the power consumption are optimized for each MDAC stage.

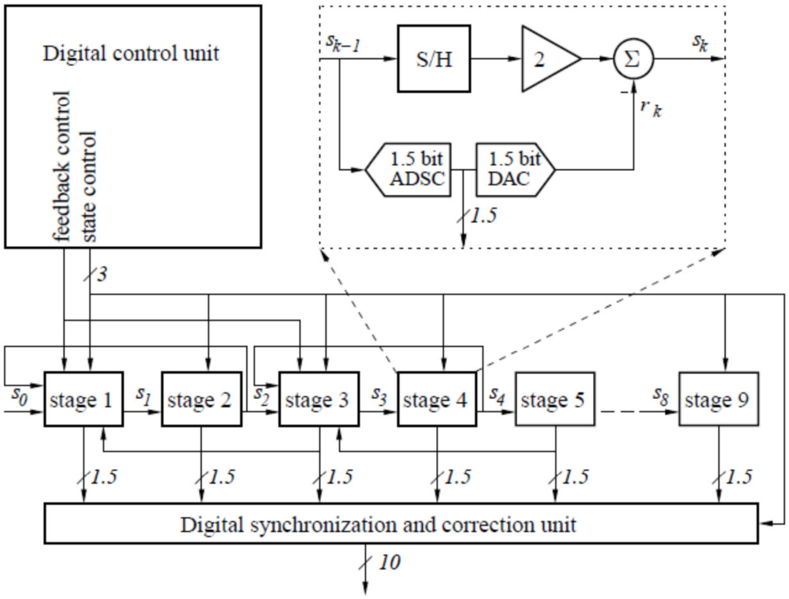


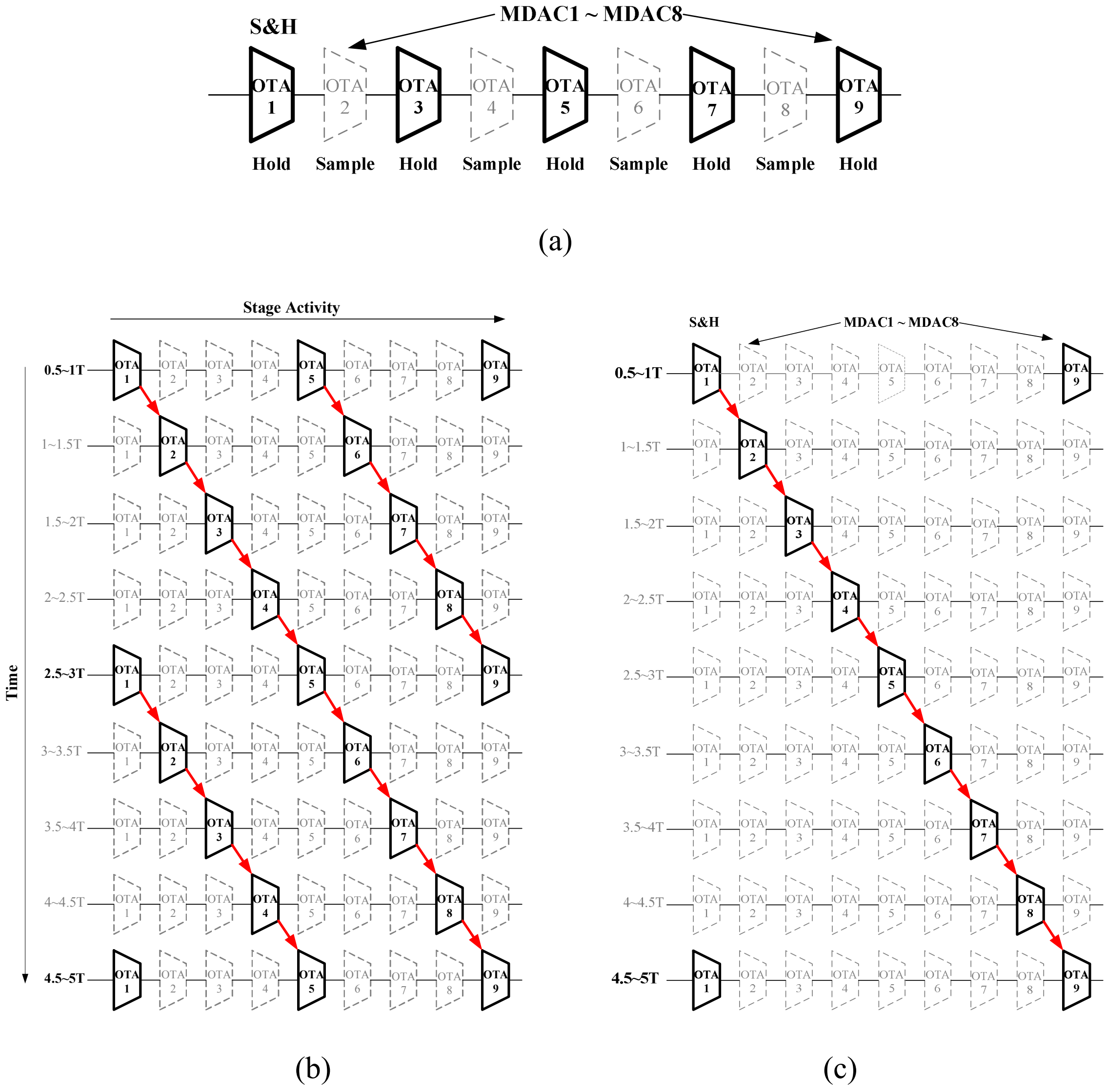
Fig. 6.3. A hybrid pipeline/cyclic reconfigurable ADC [96] (© 2005 IEEE).

The effective speed can be reduced by configuring a pipeline ADC in cyclic mode. [96] reported a hybrid pipeline/cyclic reconfigurable ADC, as illustrated in Fig. 6.3. Two residue feedback loops are introduced to operate stage1/stage2 and stage3/stage4 as cyclic ADCs during certain clock cycles. However, amplifiers in the MDAC stages see different loadings when driving succeeding/preceding stages, and stage scaling cannot be efficiently applied to optimize power. Therefore a better reconfigure method is needed to fully leverage the potential of pipeline and cyclic ADC.

6.3 Proposed Reconfiguration Architecture

6.3.1 Proposed "Global Cyclic" Technique

Based on the above observations, we propose the "Global Cyclic" technique for implementing a reconfigurable ADC. Fig. 6.4 shows the conceptual diagram. It is based on a 10bit pipeline ADC with a S&H, eight 1.5bit MDAC stages and a 2bit flash. Notations for the S&H and MDACs have been simplified by means of an OTA, while switches and capacitors are not included. Thus there are nine OTAs involved, each represented by a trapezoid. The solid-line trapezoid means that the stage is in hold mode, while the dashed-line means it is turned off during the sampling phase. Fig. 6.4 (a) shows the ADC configuration in full speed mode Fs (i.e. the input is sampled every T, T=1/Fs), where it works as a typical pipeline and to save power, the OTA is only powered on in the hold phase, thus the average power is: 5* ½ + 4* ½ = 4.5. Fig. 6.4 (b) shows the ADC configuration at Fs/2, i.e. the input is sampled every 2T (i.e. at time instant 0.5T, 2.5T, 4.5T). Note that there is only one physical row, but we are expanding it in time (the vertical axis) to show the operation more clearly. The arrows represent the track of an analog input. Shortly before one stage powers off, the subsequent stage



(a)

(b)

(c)

Fig. 6.4. Proposed “global cyclic” technique: (a) full speed mode (Fs) (b) Fs/2 speed mode (c) Fs/4 speed mode.

powers on and enters the hold mode to ensure the input continuously goes through the entire pipeline chain. The digital outputs from each stage are latched before that stage powers off. The average power is: 3* 1/4 + 2* 3/4 = 2.25. Fig. 6.4 (c) shows the Fs/4 mode, where the input is sampled every 4T with the average power as: 2*1/8 + 1*7/8 = 1.125. In this mode, it essentially operates as a cyclic ADC because only one stage is working at a time. It can be viewed as unfolding a cyclic ADC in space along a pipeline chain. Table 6.2 illustrates the power scaling for the proposed "Global Cyclic" technique. Theoretically the power scales at the same ratio as the speed scales, keeping a constant power/speed ratio and FOM.

Table 6.2. Power scaling for the "global cyclic" technique

Speed	N	Sampling Interval	Average Power	Normalized Power
Fs	1	T	5* 1/2 + 4* 1/2 = 4.5	1
Fs/2	2	2T	3* 1/4 + 2* 1/4 = 2.25	1/2
Fs/4	4	4T	2* 1/8 + 1* 7/8 = 1.125	1/4
Fs/8	8	8T	2* 1/16 + 1* 7/16 = 0.5625	1/8
……				
Fs/N		NT	2* 1/2N + 1* 7/2N = 9/2N	1/N

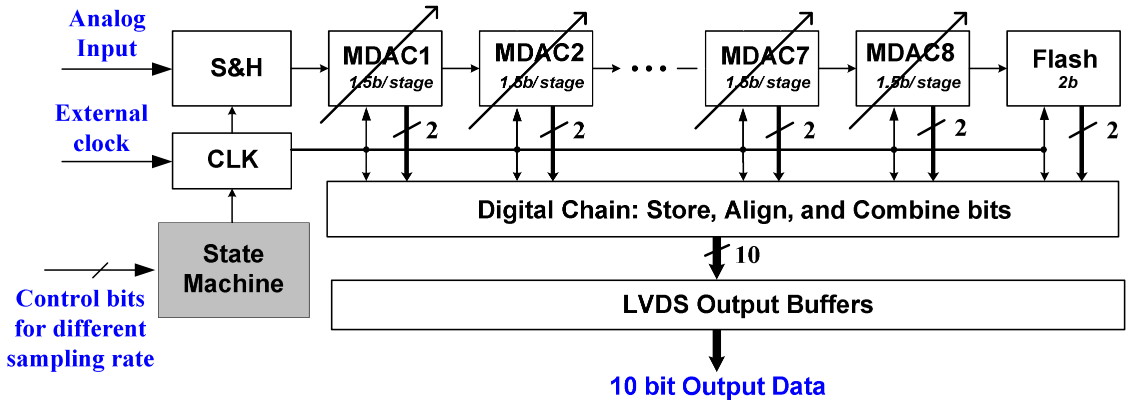


Fig. 6.5. System diagram of the proposed reconfigurable ADC.

Fig. 6.5 shows the block diagram of the proposed reconfigurable ADC. It has two unique features: 1) a state machine is added to generate the power on/off timing signals according to external control bits to achieve different effective sampling rates. 2) The duty cycle of each MDAC stage is programmable. While a typical cyclic ADC circulates the residue signal, the proposed ADC performs “pseudo circulation”: the residue signal still passes from one stage to the next, but the power consumption averaged over a conversion cycle. The capacitances and bias currents in MDAC2-5 are scaled down by a factor of 0.55 compared to the 1st stage, and those in MDAC6-8 are further scaled down by 0.52. The stage scaling along the pipeline chain decreases the power consumption.

Fig. 6.6 depicts a comparison between the proposed “Global Cyclic” and the typical “current scaling” techniques. For the “current scaling”, the ADC is always “ON”, and power is scaled by adjusting the sampling clock period and the bias current at the same ratio. For the “Global Cyclic”, the bias current is kept constant (i.e. same pulse width, same setting time/accuracy) thus the ADC performs conversions at a constant maximum rate. The effective speed is programmed by varying the ADC's "ON" time. The same averaged power consumption is achieved between these two approaches, but the "Global Cyclic" has two advantages over the "current scaling" approach: 1) only one sampling clock is required, simplifying the system requirment; 2) the bias current is kept constant, eliminating the reliability issue. The "Global Cyclic" ADC also has two advantages over the previous "pipeline/cyclic reconfiguration [96]": 1) it combines well with pipeline stage scaling; 2) apart from the state machine, no extra digital logic is needed for the cyclic mode.

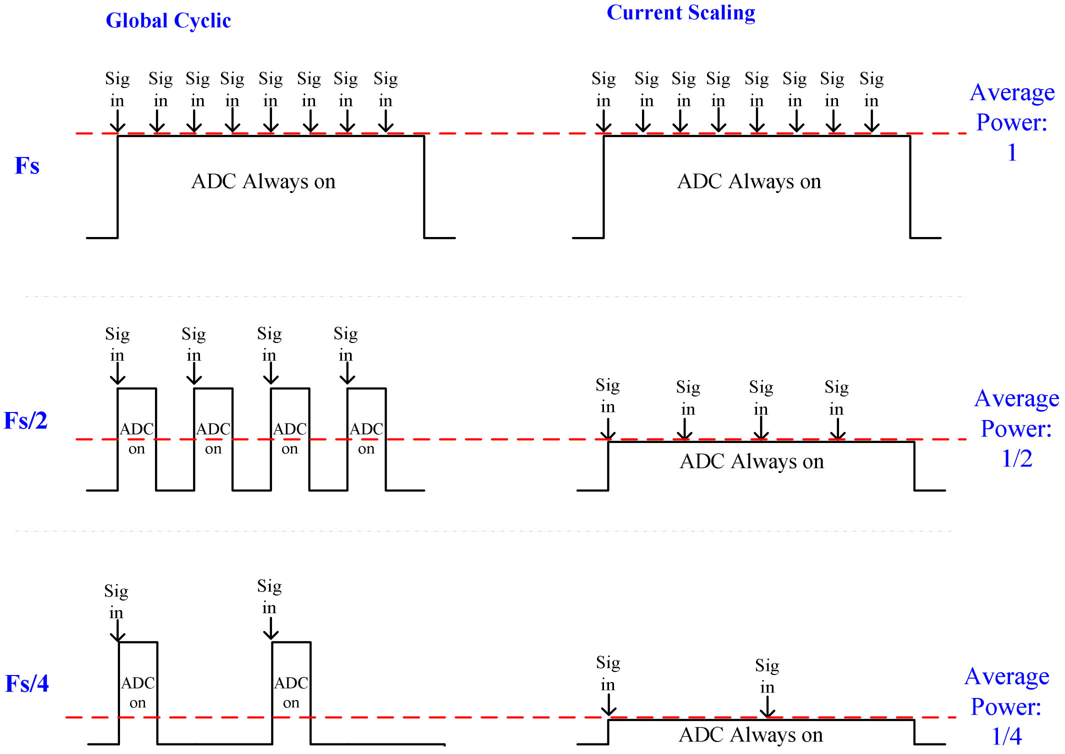
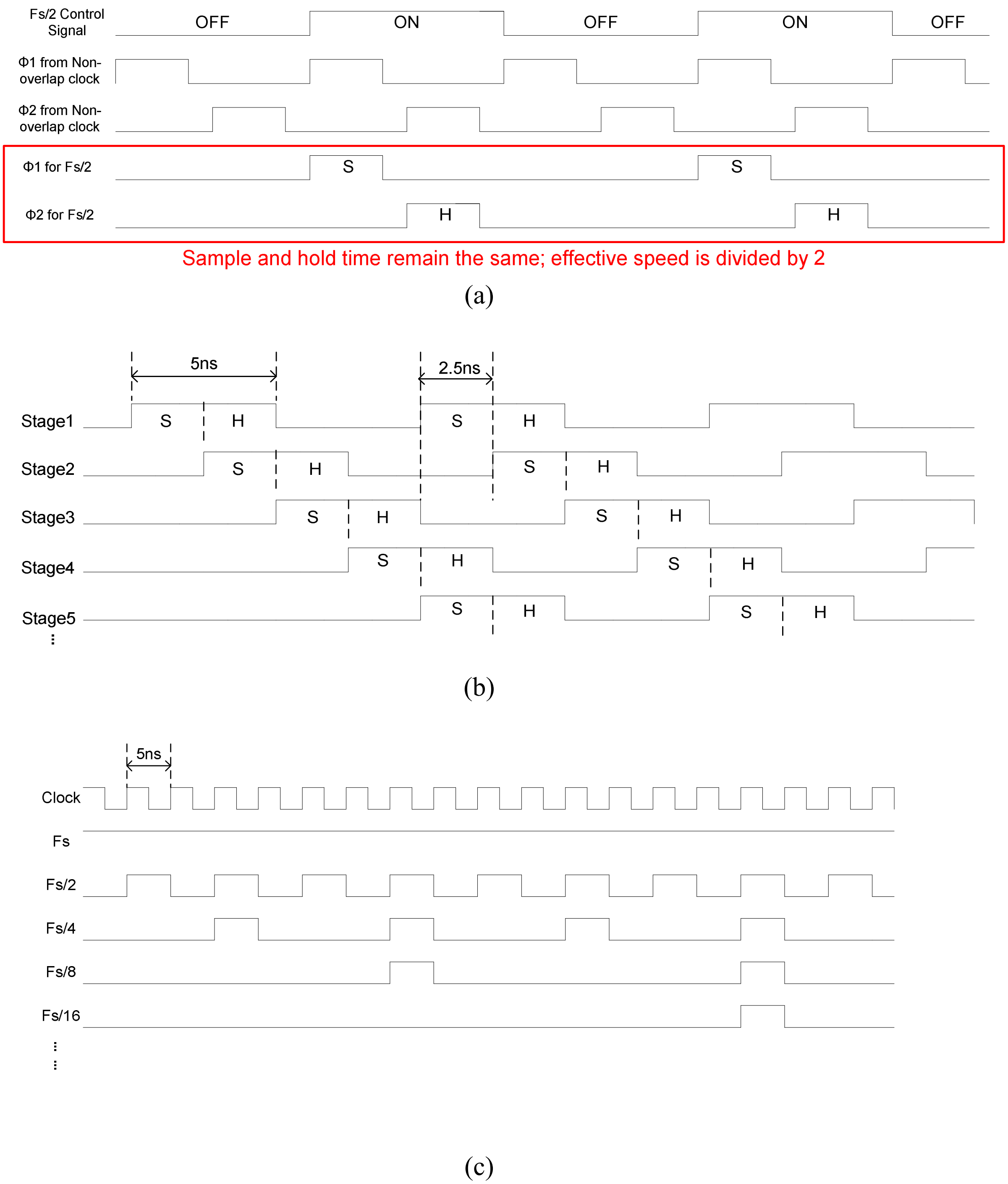


Fig. 6.6. "Global cyclic" vs. current scaling.

6.3.2 State Machine

The state machine is one of the key blocks to achieve reconfigurable speed and scalable power, and it generates different control signals according to various sampling rate requirements, i.e. Fs, Fs/2, Fs/4…Fs/256. The control signals are fed into "AND" gates together with the non-overlapping clock generator's outputs to generate the actual clocking control signals for the switches in each stage. Fig. 6.7 (a) shows an example of how the control signals from the state machine and the clock signals from non-overlapping clock generator generate the actual clocking signals for Fs/2. Fig. 6.7(b) shows the state machine control signals for stages 1-10 at Fs/2, and Fig. 6.7 (c) shows the state machine control signals for *one stage* at various sampling rates. Note that there is a larger latency (largest for Fs/256) for the control signal at lower effective sampling rate.

One challenge of achieving good power scalability is that a portion of the digital control logic is always kept active, which ultimately limits further scaling down of power consumption [91].



Sample and hold time remain the same; effective speed is divided by 2

(a)

(b)

(c)

Fig. 6.7. State machine output: (a) sampling rate: Fs/2 (b) control signal for stage 1-10 @ Fs/2 (c) control signal for various sampling rates (Fs, Fs/2, Fs/4, etc) for one stage.

6.4 Building Block Design

On the circuit level, the ADC is implemented with fully differential switched capacitor blocks. A flip-around S&H stage is used, followed by eight 1.5bit MDAC stages and a 2bit flash as the last stage. Fig. 6.8 and Fig. 6.9 show the schematic of the S&H and MDAC stage respectively. Metal-Oxide-Metal (MOM) capacitors, which are standard in the logic process with much lower cost than Metal-Metal (MIM) capacitors, are use as the sample and hold capacitors. The values for S&H, MDAC1, MDAC2~5, and MDAC 6~8 capacitors are chosen as 900fF, 500fF, 250fF, and 130fF, respectively, according to matching accuracy requirements.

6.4.1 Fast Switched OTA

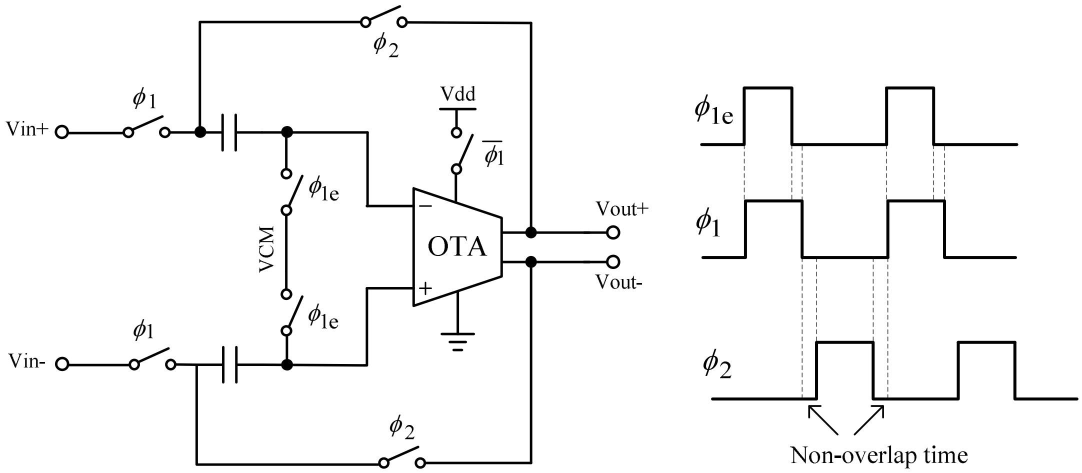


Fig. 6.8. Flip-around S&H stage.

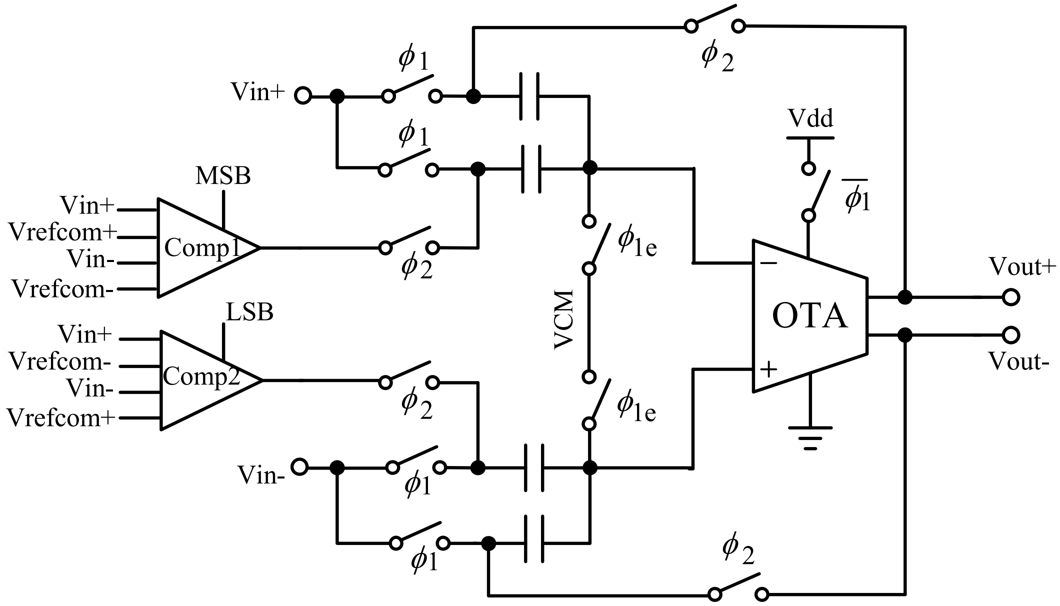


Fig. 6.9. 1.5bit MDAC stage.

The main power consumption of the ADC is from the OTAs in the S&H and MDAC stages. The OTA design is based on the recycling folded cascode architecture [134] shown in Fig. 6.10. A PMOS input folded cascode and an NMOS input folded cascode are used as the Nbooster and Pbooster, respectively. The same OTA architecture is used in the S&H and MDAC1-5 stages, but with scaled-down bias currents. For MDAC6-8, gain-boosters are removed due to the relaxed OTA DC gain requirements.

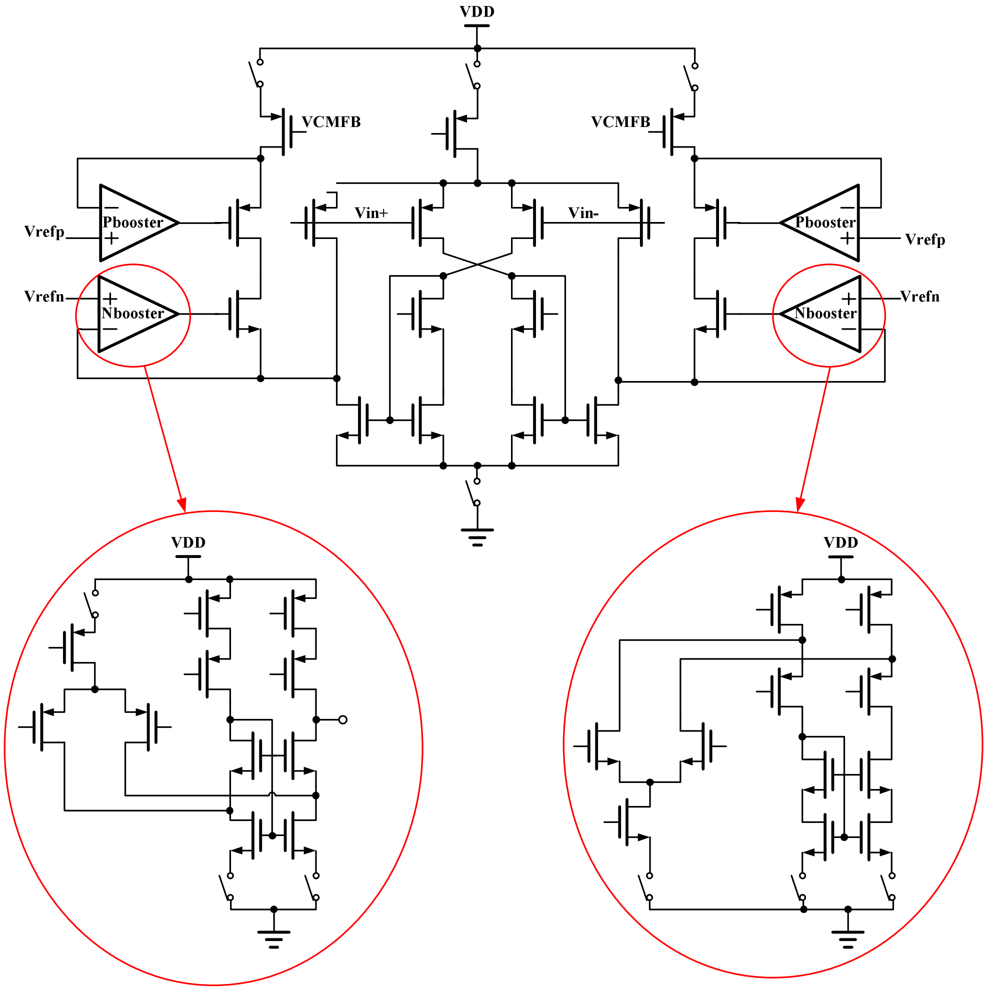


Fig. 6.10 Rapid power-on gain-boosted recycling folded cascode OTA.

The most challenging requirement for the OTA is the power-up time. At 200MSPS, the hold phase is around 2.3ns (taking into account the margin for the non-overlap time). Therefore, we have set a 200ps "lead time" to assist the OTA settling, i.e. the OTA is powered up 200ps earlier than the start of the hold phase. In our post-layout simulation, this lead time is sufficient to guarantee no degradation in the ADC resolution due to OTA settling. Switches in the current paths are added to turn the OTA on/off, and large switch sizes are used to minimize degradation in the signal swing.

A critical design issue is the settling of common-mode feedback (CMFB) circuits. Fig. 6.11 shows the switched-capacitor CMFB circuit. In phase 1, the OTA is turned off, and the outputs are reset to the desired output common voltage level VCMout. The two CMFB capacitors are reset to have a voltage across them equal to (VCMout – VbpCMFB), where VbpCMFB is the desired biasing voltage for the PMOS current source. In phase 2, these two caps are directly connected between the OTA outputs and the gates of the PMOS current sources, and the common-mode level is setup quickly.

6.4.2 Dynamic Comparator

In each 1.5 bit MDAC stage, two comparators and some combinational logic are employed to select the proper reference level, shown as Comp1 and Comp2 in Fig. 6.9. The comparator outputs have to be valid before the following MDAC stage enters the hold phase, which is essential for generating the correct residue signal. This sets the comparator speed requirement. Note that the comparator speed is highly affected by the input voltage difference.

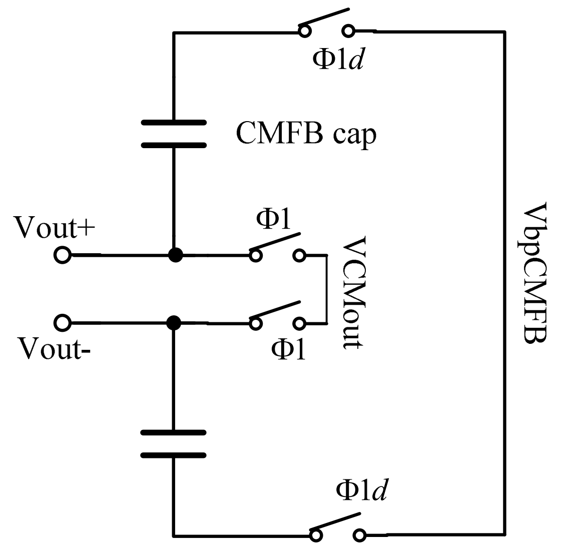
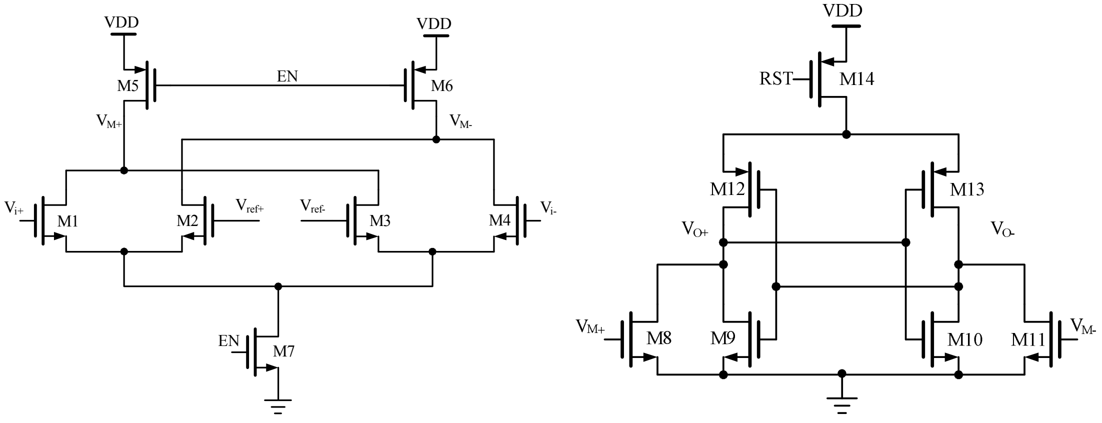


Fig. 6.11. Switched-capacitor common-mode feedback circuit.



(a)

(b)

Fig. 6.12. Dynamic comparator: (a) differential input stage; (b) latch stage.

Dynamic comparators are chosen in this design because of lower power and better power scalability. Fig. 6.12 shows the schematic. It is based on [135], but we have applied two offset reduction enhancements to it.

First, because the comparator's threshold is very sensitive to the output load capacitance due to the latch stage, dummy transistors have been added at the latch stage outputs to balance the loading capacitance.

Another important contributor of the offset is the input differential pairs. As shown in Fig. 6.12 (a), in the vicinity of the comparator threshold, the common-mode of input pair 1 (M1 and M2), $V_{ref+,}$ is higher than that of the input pair 2 (M3 and M4), V_{ref-}. In the comparison phase, as M7 is fully turned on, the input stage acts as a pseudo differential amplifier; and its transconductance is significantly affected by the input common-mode voltage level. Therefore the common-mode voltage difference between these two input pairs causes a difference in their transconductance, which results in the offset issue as derived in equation (6.4):

$$V_{out} = (V_{i,cm} - V_{ref,cm})(g_{m1} - g_{m2}) + (g_{m1} + g_{m2})(V_{\Delta i} - V_{\Delta ref}) \tag{6.4}$$

where V_{out} denotes the difference between V_{M+} and V_{M-}; $V_{i,cm}$ and $V_{ref,cm}$ are the input common-mode voltage and reference common-mode voltage, respectively; $V_{\Delta i}$ is the input differential mode voltage and $V_{\Delta ref}$ is difference between V_{ref+} and V_{ref-}; g_{m1} is the transconductance of M1 and M2, while g_{m2} is the transconductance of M3 and M4 in the comparison phase.

By observing the first term on the right of equation (6.4), notice that the differential output of the input stage is very sensitive to the input common-mode if we have a reasonable amount of transconductance unbalance between pair 1 and 2, which is actually the case in our design. To alleviate the transconductance unbalance between pair 1 and 2, we have chosen a wider size for pair 2 than for pair 1 to compensate its lower common-mode voltage.

6.4.3 Input Clock Buffer

The most important requirement for the ADC clock is low jitter. As shown in equation (6.5) [136], for an SNR > 60dB at 100MHz input, the RMS jitter of the clock should be < 2ps. It is also desired to have a 50% duty cycle for an optimal design.

$$SNR_{jitter} = -20 \cdot \log 10\left(2\pi f_{in} \sigma_{jitter}\right) \tag{6.5}$$

where f_{in} is the analog input signal frequency, and σ_{jitter} is the rms jitter of the sampling clock.

To have better signal integrity on the PCB and to minimize jitter, we feed an off-chip differential low swing sine wave to the ADC clock input pin and use an on-chip clock buffer to convert it into a single-ended square wave clock. As shown in Fig. 6.13, a simple differential pair performs the differential-to-single-ended conversion; the output is then gained up by two inverters and fed into a divide-by-two circuit to obtain an accurate 50% duty cycle. The penalty is that the input signal has to be twice the frequency of the sampling clock.

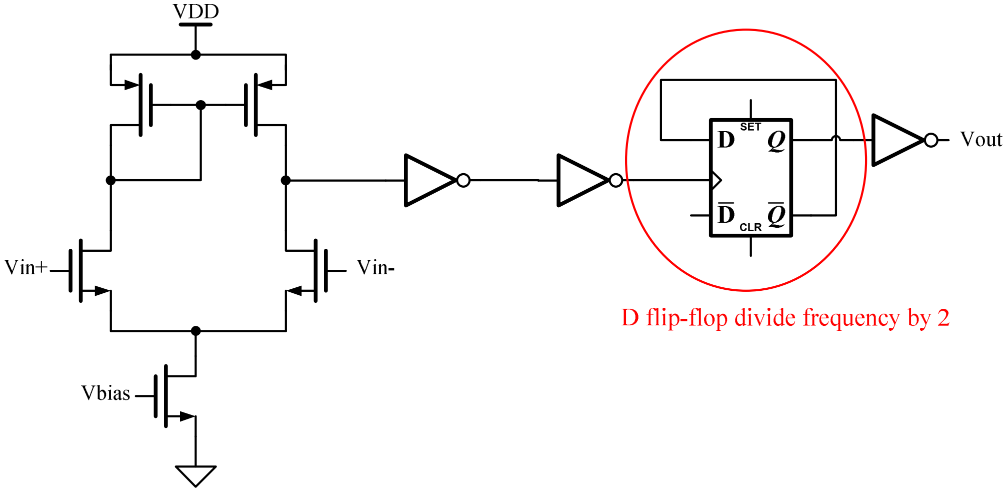
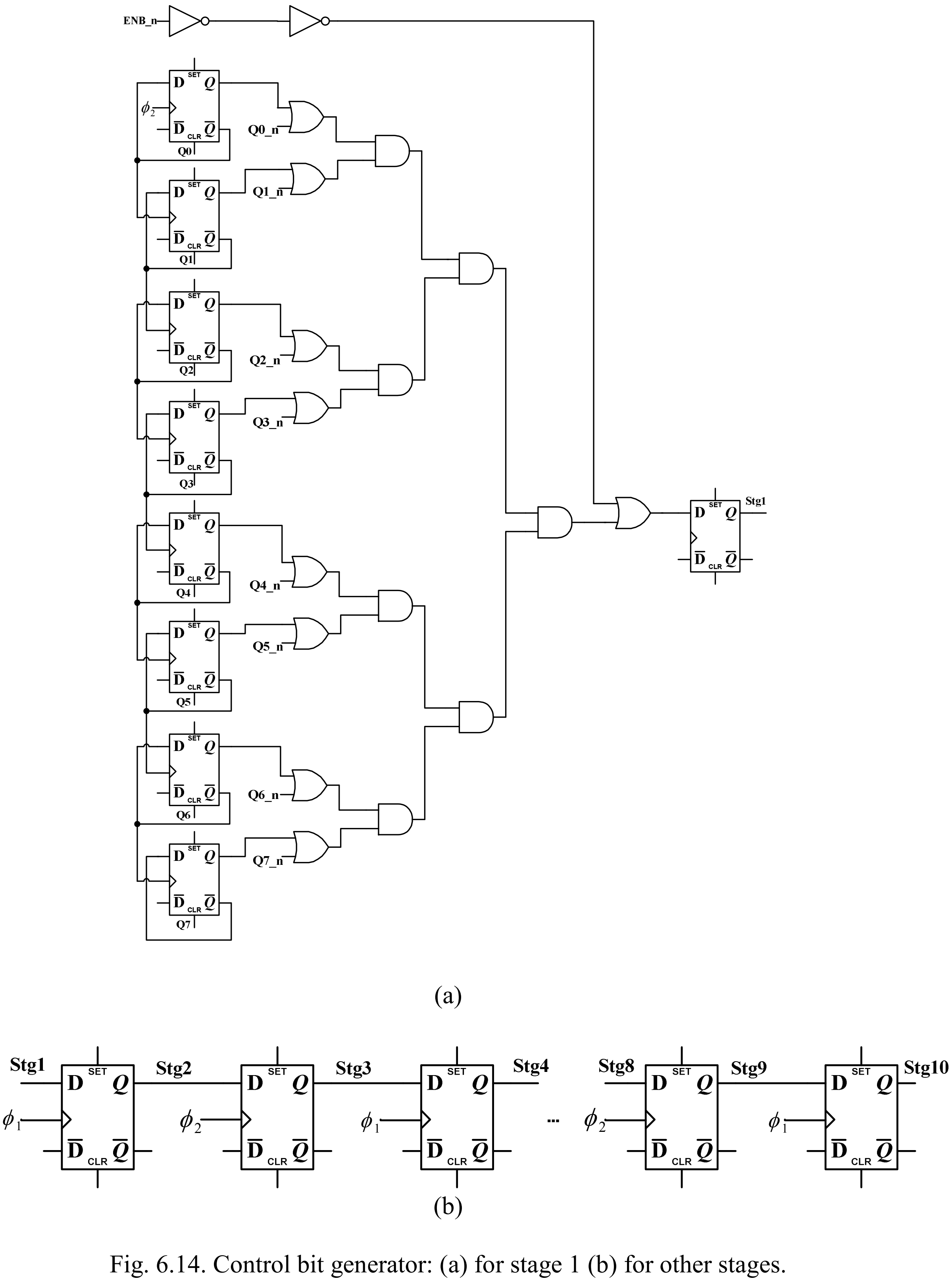


Fig. 6.13. Input clock buffer.

6.4.4 Digital Logic

As shown in Fig. 6.5, a state machine generates the signals for controlling the duty cycles of each MDAC stage. Fig. 6.14 (a) shows how the control signal of the first stage is generated. An 8-bit asynchronous counter and a three-to-eight decoder are designed to configure the sampling rate. The counter is composed of eight falling edge-triggered D flip-flops (DFFs). Phase Φ2 from the non-overlapping clock generator is used as the clock for the first DFF. For each DFF, output $\bar{Q}$ connects to its own input D and also to the clock of its following DFF. The output of the three-to-eight decoder controls eight DFFs to select different sampling rate. For example, all of the eight DFFs are disabled for the sampling rate of Fs. For Fs/2, only the first DFF is enabled and the other seven DFFs are disabled. For Fs/4, the first two DFFs are enabled and the other six DFFs are disabled, and so on. Notice that the control signal has an variable duty cycle, which is 50% for Fs/2, 25% for Fs/4, 12.5% for Fs/8, etc. Also, the delay of the circuit is



(a)

(b)

Fig. 6.14. Control bit generator: (a) for stage 1 (b) for other stages.

determined only by the first DFF and its following logic gates since the output only changes when the first DFF output changes from 0 to 1. Thus, the delay time is fixed during different sample rates.

Fig. 6.14 (b) shows the control signal generator for stages 2-10. The signal of each stage is delayed by Ts/2 compared to its previous stage. For each stage, there is one falling-edge-triggerred DFF whose input is the control signal of the previous stage. The hold phase clock of each stage serves as the clock for the corresponding DFF.

6.4.5 Power Scalability

Both the state machine and clock provider power scale with sampling rate. However, for the control clock generator, power doesn't scale at the same rate as the frequency, this is because more DFFs and logic gates are enabled at lower sampling rate; but meanwhile, the operating speed of each gate is decreasing together with frequency.

For the analog part, since the main power consumption comes from the OTAs whose duty cycle scales with sampling rate, the power scales well with speed. Other than that, both the dynamic comparator and sampling network have very good power scalability.

6.5 Conclusions

This chapter introduced a "Global Cyclic" scheme for an efficient implementation of speed programmable/power scalable ADC working at medium-to-high speed range. The work presented four main contributions: 1. good power scalability: the power consumption scales linearly with the sampling rate at xmW/MSPS in the entire speed programming range; 2. Robust performance: the bias current is kept constant, and transistors are in optimum operation region for the whole programming range; 3. Comparable FOM with state-of-art dedicated ADCs with similar specs; 4. Wide programming speed from 0.8MSPS up to 200MSPS covers all video formats and is well suited for a wide range of applications.

CHAPTER VII

LAYOUT CONSIDERATION AND EXPERIEMENTAL RESULTS FOR THE RECONFIGURABLE ADC

7.1 ADC Layout Design and Consideration

The ADC was laid out by Virtuso layout editor from Cadence. The layout extraction, DRC and LVS check were performed using Assura in Cadence. The ADC was designed and fabricated in UMC 90nm Logic/Mixed mode CMOS process. Fig. 7.1 shows the chip micrograph. The active area occupies 1.6mm by 0.95mm.

The digital circuit noise is one of the major sources of ADC performance degradation. To avoid the digital noise coupling, all the noisy digital logic, clock buffers, and output buffers are located in the top half of the chip, and surrounded by p+ substrate contacts and the n-well guard ring; while the analog circuits are kept away from the digital circuits by sitting at the bottom half of the chip and surrounded by guard rings. From the left to the right of the analog chain, as marked in Fig. 7.1, is SOH, MDAC 1-8 and 2bit flash. It has been arranged in this way to ensure shortest distance from the input clock to SOH to minimize jitter. For the clock distribution in a pipeline-based ADC topology, we just need to guarantee proper non-overlap time between consecutive stages after taking into the routing parasitic, therefore the clock generator is placed on the left corner, and the clock signals are distributed from left to right, instead of using any advanced layout techniques such as “H-tree”.

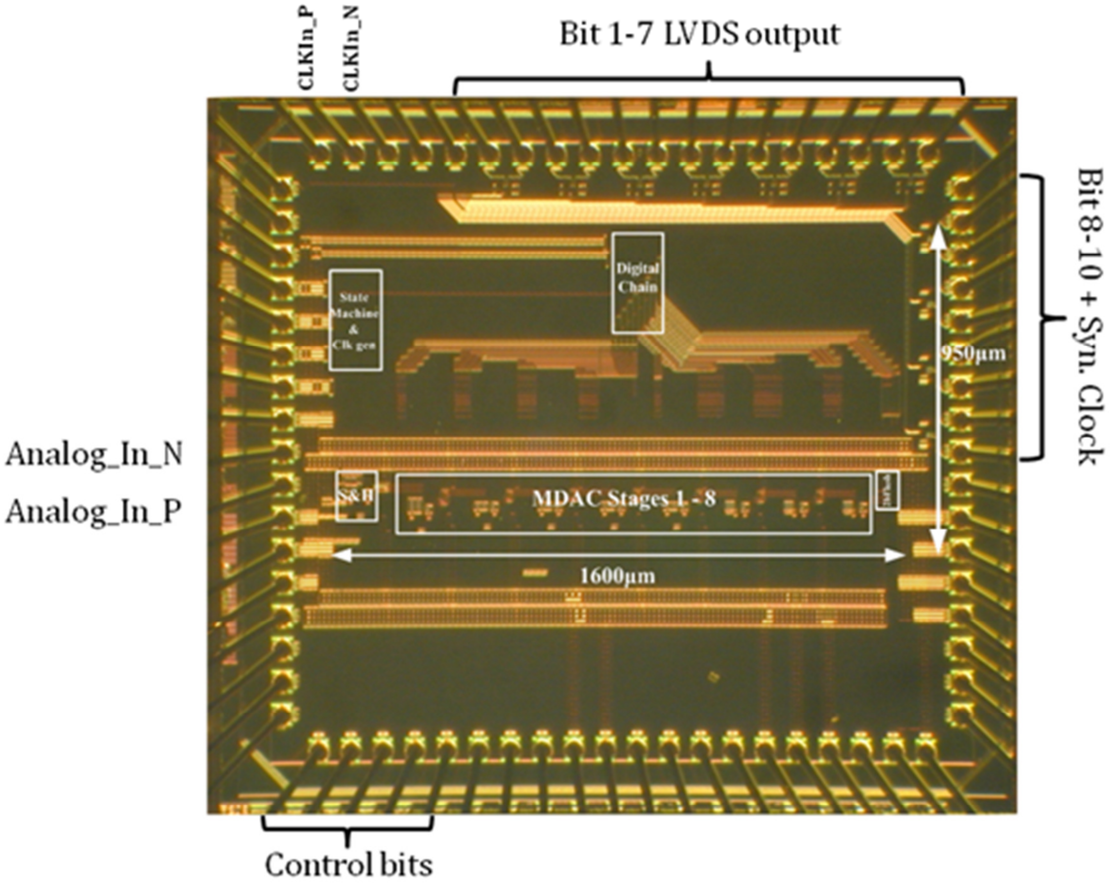


Fig. 7.1. Chip micrograph of the proposed reconfigurable ADC.

The fully differential analog input signal is applied to the ADC from the middle of the left side of the chip for best symmetric when considering bond wire effects. It is critical to arrange the analog input and clock input to be orthogonal with sufficient isolation pads (i.e. DC pads) in between to minimize the coupling. Thus the fully differential 400 MHz clock is applied from the top side of the chip. Although putting it in the middle of the top edge could have better symmetric, we decided to apply the clock from the top left corner in order to minimize the routing to the non-overlap clock generator for better jitter performance. For better signal integrity at 100MHz, we have used low swing differential signaling (LVDS) output buffers to take the 10 bit digital signal and the synchronizing clock off chip.

A few pads have been dedicated for power supplies and grounds, which are distributed around the chip to minimize the IR drop.

7.2 Print Circuit Board (PCB) Design

Fig. 7.2 shows the FR-4 PCB for ADC testing. The ADC chip, placed in the middle of the PCB, has been packaged in QFN 64. Sufficient decoupling caps with

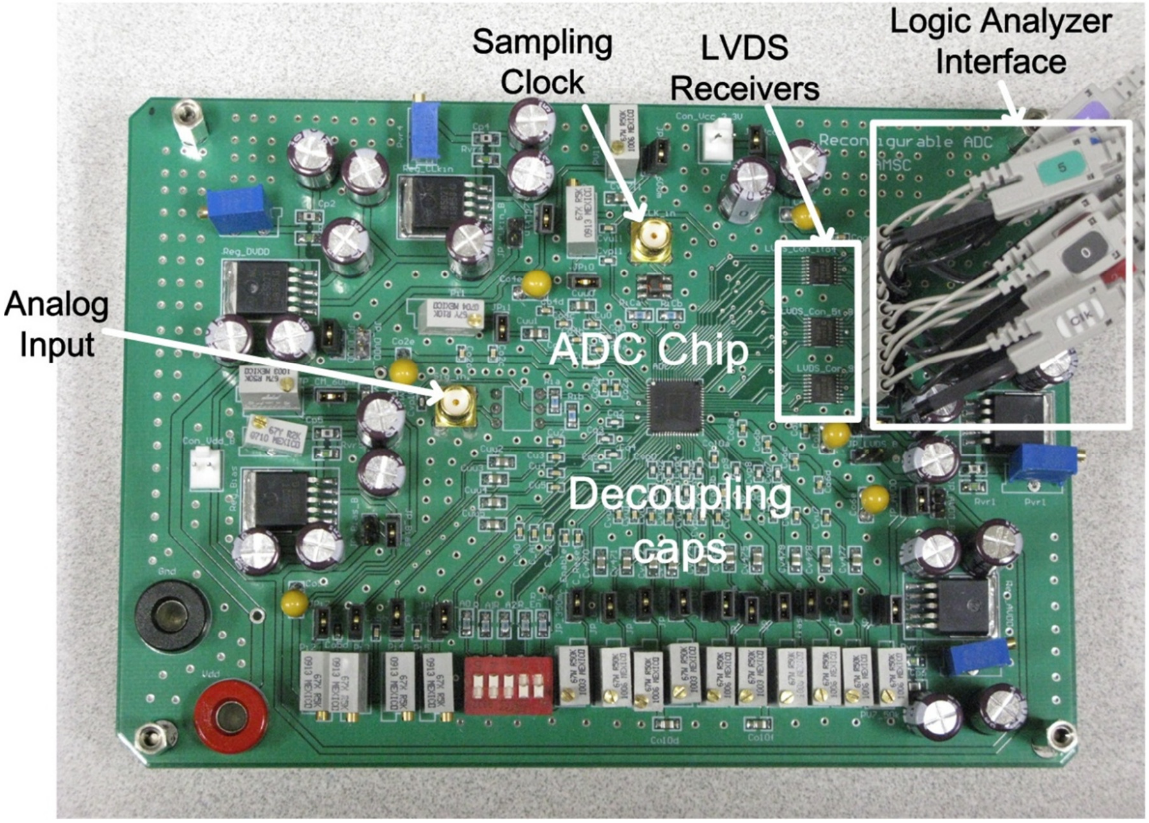


Fig. 7.2. PCB picture.

distributed self-resonant frequencies have been placed closed to the ADC chip to provide fast transient current. Regulators are used to provide a quiet and stable power supply. Three LVDS converters are placed at the ADC output to convert the low swing differential signal into single-ended TTL signal to interface with the logic analyzer.

7.3 Testing Setup

Fig. 7.3 and Fig. 7.4 shows the ADC test setup and the lab measurement picture, respectively. The signal generator and low-jitter clock generator have been synchronized at 10MHz for coherent sampling. Since the typical signal generators have harmonic distortion as high as 40dBc, for 10bit ADC testing, an external band pass filter(BPF) with at least 25dB attenuation is necessary to filter out the harmonics from the signal generator to provide a pure analog input signal for the ADC. Here a passive LC-BPF is employed. The logic analyzer captures the 10bit digital output for post-processing in PC to obtain SNDR, SFDR, DNL and INL.

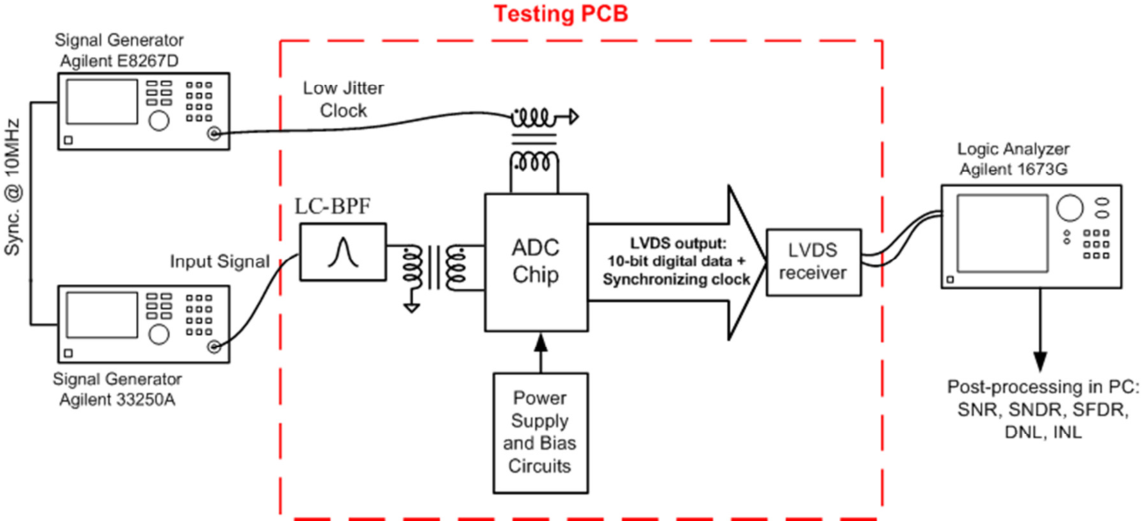


Fig. 7.3. Testing setup of the ADC.

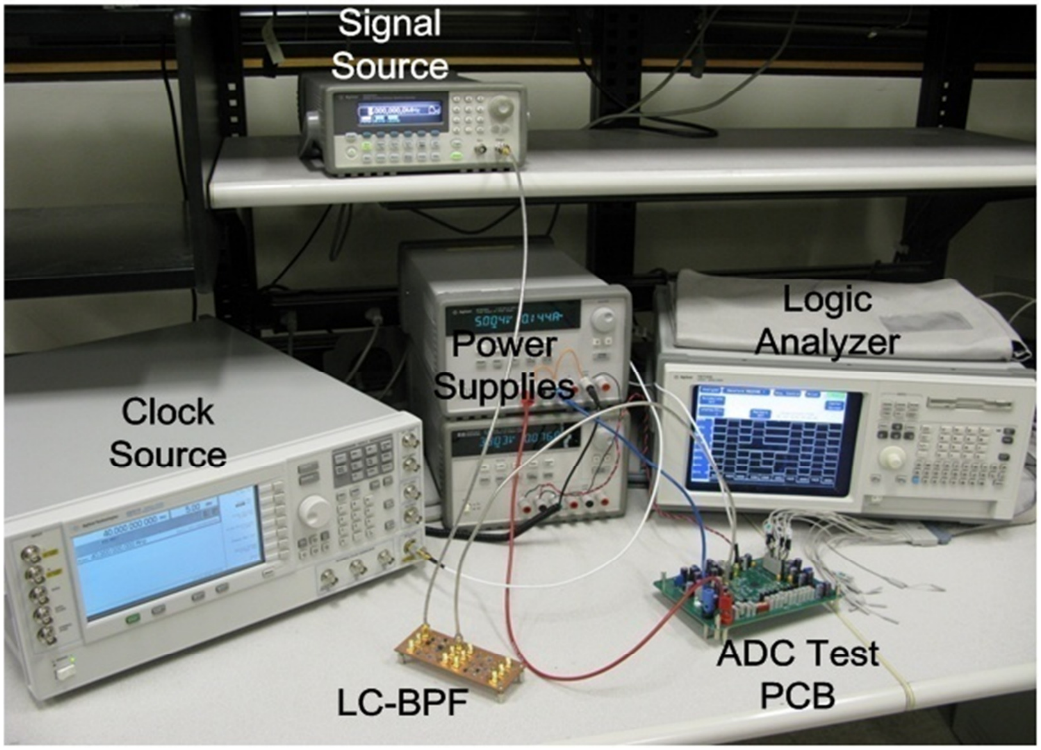
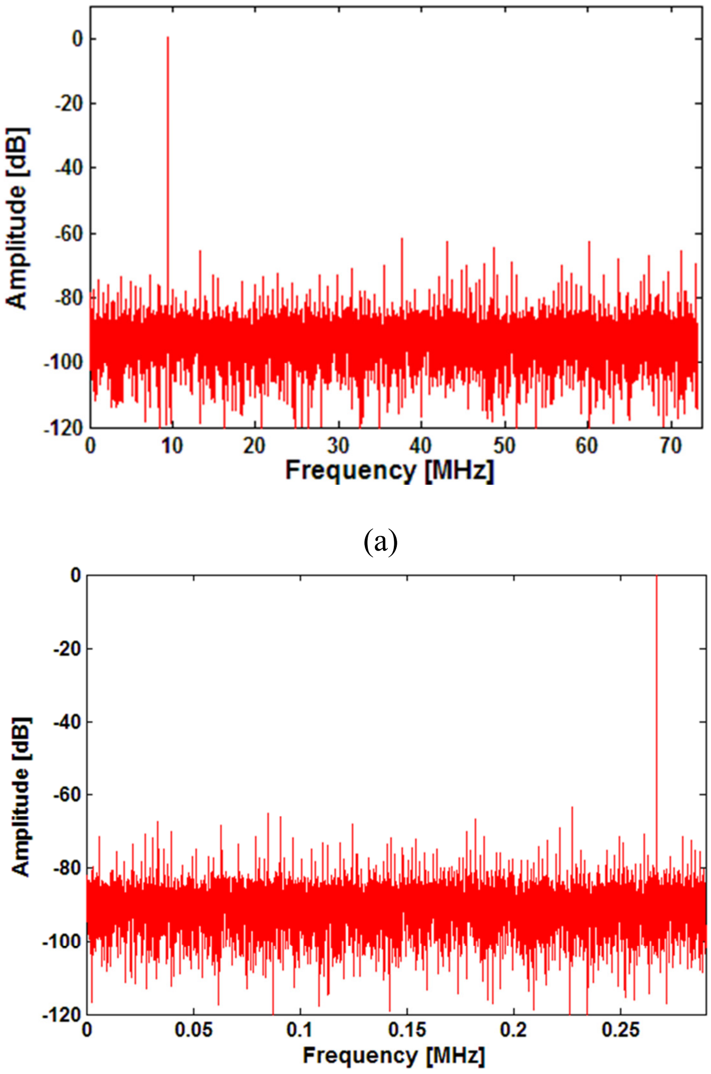


Fig. 7.4. Lab measurement picture.

7.4 Measurement Results

Both the static and dynamic performance of the ADC have been characterized. The basic metrics for dynamic performance is the signal-to-(noise+distortion) ratio(SNDR) and the spurious free dynamic range(SFDR). The basic idea is to apply one



(a)

(b)

Fig. 7.5. ADC output spectrum (a) Fin = 9.4MHz, Fs = 150MHz. SNDR = 52dB, SFDR = 63.1dB (b) Fin= 0.26MHz, Fs/256 = 0.58MHz. SNDR = 51dB, SFDR = 64dB.

tone at ADC input, and we expect the same tone at the output, while all other frequency components represent non-idealities. Fig. 7.5 shows the measured frequency spectrum of the ADC at Fs = 150MHz and Fs/256 = 0.58MHz, which are the two extremes in the ADC speed reconfigurable range. Fig. 7.6 shows the measured SNDR and SFDR of the ADC versus the input frequency at 150MSPS. SNDR is above 49dB up to the Nyquist frequency; while SFDR is above 59dB over the full Nyquist band. The SNDR and SFDR are also plotted as a function of the sampling rate as shown in Fig. 7.7. SNDR varies less than 2dB within the entire speed programming range, while the variation in SFDR is kept below 3dB. This consistent performance over a wide speed range is as expected because the bias current are kept constant when we are programming the ADC speed, therefore the circuit works robust. The input signal swing is 1Vpp for these measurements at 1.1V power supply.

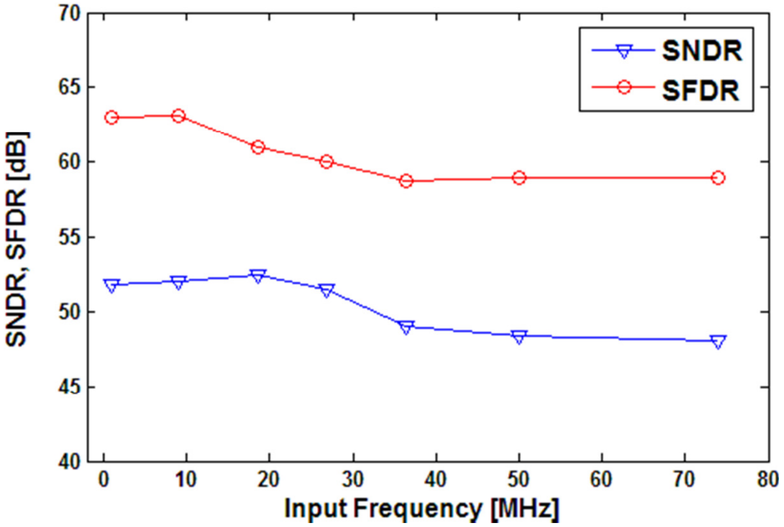


Fig. 7.6 SNDR, SFDR vs. input frequency @ 150MSPS.

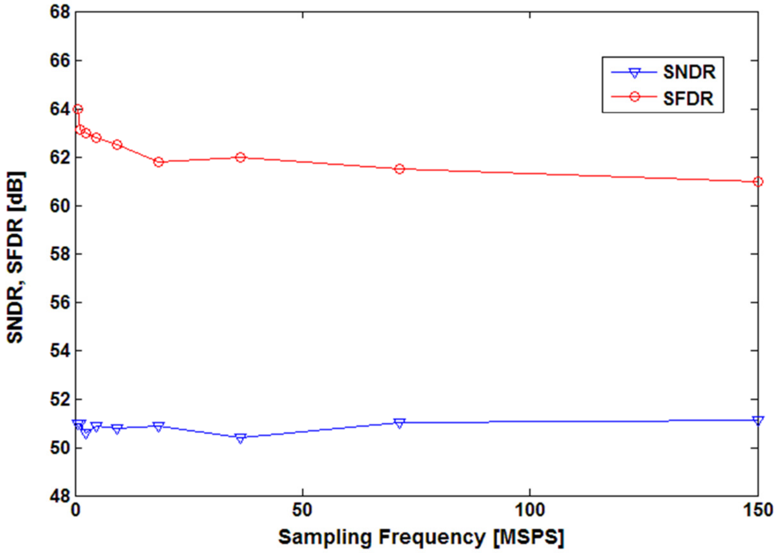


Fig. 7.7. SNDR, SFDR vs. sampling frequency.

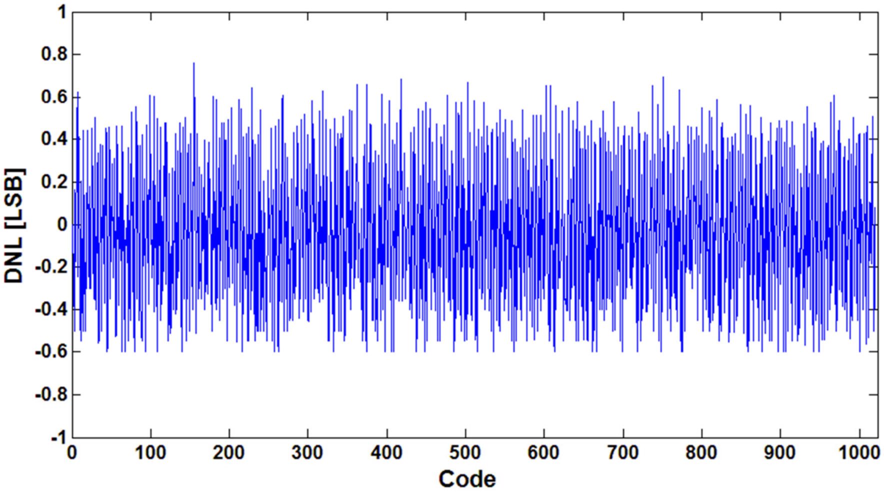


Fig. 7.8. DNL @ 150MSPS: -0.6/+0.76LSB.

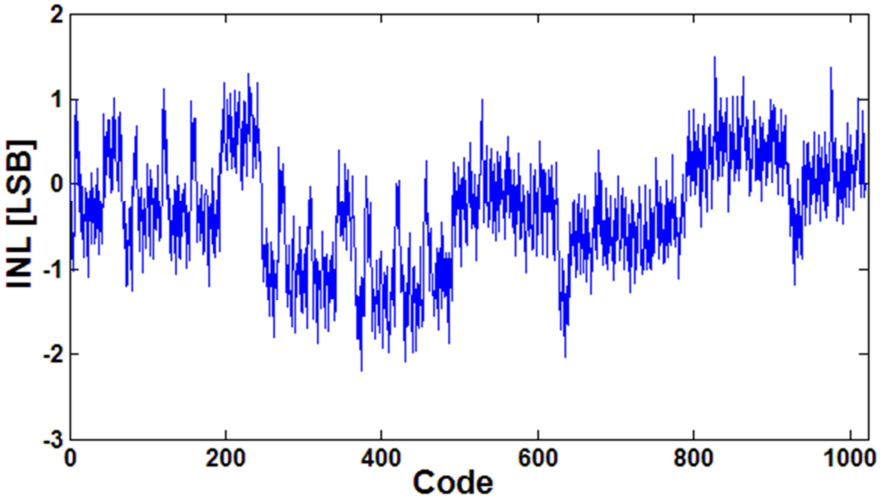


Fig. 7.9. INL @ 150MSPS: -2.1/+1.5LSB.

Two metrics to characterize the static performance of the ADC include: 1) differential nonlinearity (DNL), which is a measure of uniformity, and ideally each code has the same width; 2) integral nonlinearity (INL), which is a measure of linearity, and the ideal transfer function is a straight line through end points. Fig. 7.8 shows the measured DNL at 150MSPS, which is less than 0.76LSB. Fig. 7.9 shows that the measured INL at 150MSPS is less than 2.1 LSB.

To show the ability of the proposed ADC to adapt its power consumption to the needed speed, the power dissipation as a function of the speed has been plotted in Fig. 7.10 where the sampling rate is swept from 0.58MSPS to 150MSPS, the power is proportional to the effective sampling frequency. Fig. 7.11 compares the power of this reconfigurable ADC with the state-of-art customized 10bit ADCs in the entire speed programmable range. Comparable power consumption has been achieved.

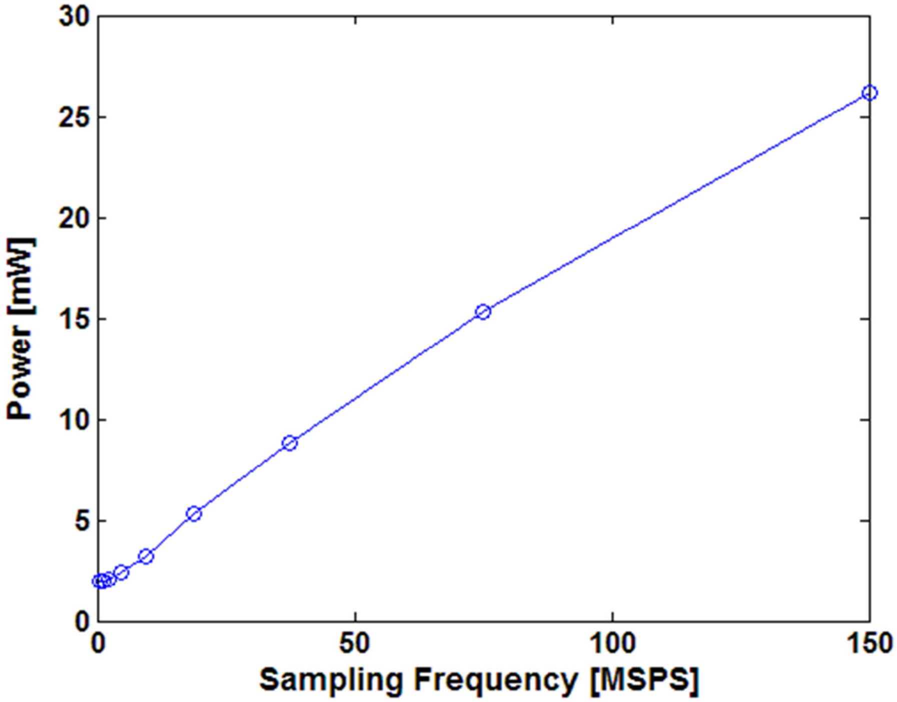


Fig. 7.10. Power vs. sampling rates.

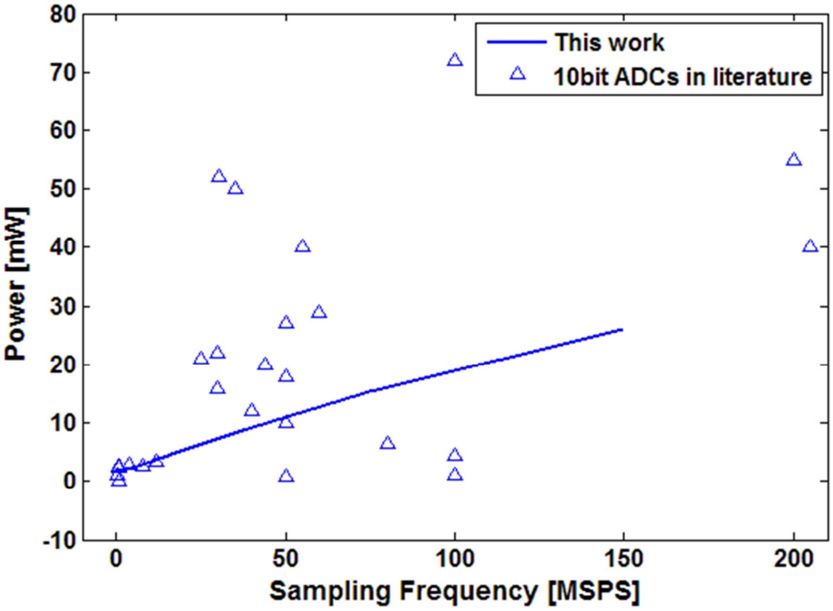


Fig. 7.11. Comparison with state-of-art customized ADCs.

A comparison of state-of-art reconfigurable ADCs is listed in Table 7.1. The current scaling technique yields good power stability, but limited speed programming range. The hybrid pipeline/cyclic approach has sub-optimal power consumption because there is no stage scaling; the time-interleave technique also has limited speed programming range, and it requires complex clock distribution and involves mismatch problem for increased number of parallel branches. The proposed Global Cyclic technique achieves a wide speed program ratio with the highest sampling rate.

Table 7.1. Comparison of the state-of-art speed/power reconfigurable ADCs

	Resolution	Speed (MSPS)	Power	Process	Technique
This Work	10bit	0.58-150	1.9-27mW	90nm	Global Cyclic
G.Geelen ISSCC 06	10bit	25-120	0.3mW/Msample	90nm	Current Scaling
B. Hernes ISSCC 04	10bit	3-220	90mW@120M 135mW@220M	0.13µm	Current Scaling
M. Anderson VLSI 05	6-10bit	20/40/80	30.3/52.6/93.7mW	0.18µm	Hybrid Pipeline/Cyclic
I.Ahmed JSSC 05	10bit	0.001-50	15µW-35mW*	0.18µm	Sleep mode + Current scaling
B. Xia JSSC 06	10bit	11/44	14.8/20.2mW	0.25µm BiCMOS	Time interleave +current scaling

* The digital power is not included

CHAPTER VIII

CONCLUSIONS

8.1 Summary

This research work studied two building blocks, the LNA and the ADC, for the next generation multi-standard/wideband applications. The LNA requires broadband frequency response and high linearity, and the ADC requires reconfigurability to operate under different communication standards without significantly increasing the implementation cost. A few techniques are proposed after analyzing the pros & cons of the existing solutions.

Eight categories of CMOS LNA-linearization techniques are reviewed and the tradeoffs among linearity, power, and PVT variations are discussed. General design guidelines are provided for high-linearity LNAs.

A linearization and noise reduction technique is proposed for a differential cascode LNA. The inductor connected at the gate of the cascode transistor and the capacitive cross-coupling are strategically combined to reduce the nonlinearity and noise contributions of the cascode transistors. A test chip in TSMC 0.35µm CMOS process demonstrates a 2.35dB improvement in IIP3 and a 0.35dB reduction in NF. The LNA is also designed in UMC 0.13µm CMOS process, and the proposed technique reduces the NF from 1.55 dB to 0.95 dB in simulation, which verifies its effectiveness in the deep-submicron process.

A practical linearization technique is explored for high-frequency, wideband applications using an active nonlinear resistor. The linearization technique is applied to a

UWB LNA. Experimental validation of the linearization scheme demonstrates factor of two improvement in linearity over a broad frequency range (2.5–10 GHz). The technique furthermore obtains a robust linearity improvement over process and temperature variations. The idea was verified by three UWB LNAs designed and fabricated in UMC 90nm CMOS process. The proposed UWB LNA achieves excellent linearity with much less power than the prior published state-of-art UWB LNAs.

A global offset cancellation technique is proposed to alleviate the offset and flicker noise problems in a cyclic ADC, hence reducing its power consumption and area. The cyclic ADC, designed and fabricated in TI 0.35μm CMOS process, demonstrates a better FOM compared to the state-of-art ADCs in the same process.

A "Global Cyclic" reconfiguration scheme is proposed to program the ADC sampling rate and scale its power consumption with constant biasing current. The ADC features a wide speed programming ratio while achieving good power scalability with robust performance. Implemented in a pure digital 90nm CMOS process with nominal supply voltage at 1.2V, the ADC maintains its performance down to 1V supply at a differential signal swing close to full scale (1Vpp). The measurement result shows a 54 dB SNDR for a sampling rate ranging from 0.8 MSPS to 200 MSPS, while power scales linearly at 0.3mW/MSPS. The proposed reconfigurable ADC achieves a FOM comparable with state-of-art customized ADCs.

8.2 Future Work

8.2.1 Highly Linear Wideband LNAs in Deep-submicron CMOS Process

The emerging broadband transceivers introduce new issues for wideband LNA-linearization. IIP2 is becoming just as important as IIP3, and improving P1dB is also necessary for wideband applications to improve high-signal-handling capability. Nonlinear output conductance is a new issue in deep submicron processes, and a key challenge resides in delivering high linearity with core transistors and low supply voltage in the deep submicron processes. Linearization techniques for cancelling higher-order distortion terms (beyond 3rd order), linearizing output conductance, and improving LNA P1dB still remain open problems.

8.2.2 Reconfigurable ADCs for Emerging Applications

Recently, we are experiencing a migration from existing multi-standard transceiver to the promising “software defined radio (SDR)”. Three main differences are identified between them: 1) the number of standards integrated. The SDR offers customers an integration of much more services including cellular, cordless, satellite mobile WPAN/WLAN/WiMax, Bluetooth, UWB, GPS, DAB, DVB-T/H. etc; 2) SDR requires higher flexibility to incorporate future new standards with short development time and low cost. 3) SDR has optimal power scalability. Its power consumption is minimized at each performance node, and adapt to the environment or different Quality of Service (QoS). The transceiver, driven by the QoS, needs to be dynamically adaptive, while for the existing multi-standard receivers, configurations are switched in a static sense when switching between standards.

Compared with multi-standard transceivers, the reconfigurable ADCs for SDR should cover a larger spread of specifications with better power scalability. Table 8.1 lists the ADC requirements for some state-of-art communication standards; the ADC should meet the low-bandwidth/high-dynamic-range requirements as well as the high-bandwidth/low-dynamic-range requirements. For the SDR, a wide band of RF spectrum would be digitized, and subsequently demodulated by a digital processor. Radios should adapt to any standard and intelligently manage interference and bandwidth allocation. This poses more challenges for the design of reconfigurable ADCs.

Another new application for reconfigurable ADCs is the wireless sensor networks, which has time-varying and unpredictable performance demands and energy budget. The ADC needs to handle a variety of different signals (eg. voice, sound, image, temperature, seismic, blood pressure, heart beat, etc.) in real time. A reconfigurable ADC with multi-signal conversion capability at minimal power consumption and small area would be the ideal candidate.

Table 8.1. ADC requirements for state-of-art communication standards

Standard	Bandwidth(MHz)	Resolution(bit)
GSM/EDGE	0.2	13-14
Bluetooth	1	11-12
GPS	2	10
UMTS(WCDMA)	3.84	9-10
WLAN (802.11a/b/g/n)	20-22	8-10
WiMAX	20	8-12
UWB	500	4-5

8.2.3 Adaptive ADCs

An adaptive ADC differentiates from a reconfigurable ADC in terms of intelligence level. In a “reconfigurable ADC”, control commands are independently provided from outside, either by users or by the system which employs the ADC. The ADC just reacts to the configure command. In an “adaptive ADC”, control commands are derived from the ADC itself, i.e. the ADC automatically adjusts settings based on the statistical and spectral properties of the analog signals received, without any external intervention. Apparently, adaptive ADCs have a higher intelligence level than reconfigurable ADCs, and are more “user friendly”; however, extra energy is needed to keep the “detection” portion alive. Take an analogy from mechanical engineering: the automatic requires more parts to keep an eye on the speed and thus consumes more fuel. On the other hand, the manual uses fewer parts (hence lower costs) as the human keeps an eye on the speed, but it obviously needs constant hand and foot inputs from the driver.

The core part of the adaptive ADC is a reconfigurable ADC. A dynamic controller senses the input signal information (RMS power, BW, etc.) and sends the reconfigure command to the ADC. Minimizing the acquisition time and power/area overhead for the dynamic controller is a big challenge. Bounds must be placed on both the acceptable bandwidth and amplitude, and the ADC should be able to adjust its settings within these bounds.

8.2.4 High Speed ADCs

Wide bandwidth and high speed are the ultimate goals of the wireless/wireline communication industry. Time-interleaved ADCs with low to medium resolution is an attractive solution for new applications which require high sampling rate but relaxed resolution, such as UWB systems, and wired transceiver at data rates of 20Gb/s and beyond. Furthermore, the advancement in CMOS technology and digital calibration schemes has made the development of high resolution, high speed ADC viable, which will eventually lead to the implementation of the "software-defined-radio".

REFERENCES

[1] A. Emira, A. Valdes-Garcia, Xia Bo, A.N. Mohieldin, A.Y. Valero-Lopez, S.T. Moon, C. Xin, and E. Sánchez-Sinencio, "Chameleon: A dual mode 802.11b/Bluetooth receiver system design," *IEEE Transaction on Circuits and Systems I: Regular Papers*, vol. 53, no. 5, pp. 992-1003, May 2006.

[2] E. Sánchez-Sinencio and J. Silva-Martinez, "CMOS transconductance amplifiers, architectures and active filters: A tutorial," *IEE Proc. –Circuits Devices Syst.*, vol. 147, no.1, pp. 3-12, Feb. 2000.

[3] V. Aparin, "State-of-the-art techniques for high linearity CMOS low noise amplifiers," *IEEE RFIC Symposium Workshop WSC*, June 2007.

[4] B. H. Leung*, VLSI for Wireless Communication*. Englewood Cliffs, NJ: Prentice-Hall, 2002.

[5] W. Sansen, "Distortion in elementary transistor circuits," *IEEE Trans. Circuits Syst. II*, vol. 46, no.3, pp. 315-325, Mar. 1999.

[6] B. Toole, C. Plett, and M. Cloutier, "RF circuit implications of moderate inversion enhanced linear region in MOSFETs," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 51, no. 2, pp. 319–328, Feb. 2004.

[7] T. H. Lee*, The Design of CMOS Radio-Frequency Integrated Circuits*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[8] S. Narayanan, "Application of Volterra series to intermodulation distortion analysis of transistor feedback amplifies," *IEEE Trans. Circuit Theory*, vol. 17, no. 4, pp. 518-527, Nov. 1970.

[9] V. Aparin and C. Persico, “Effect of out-of-band terminations on intermodulation distortion in common-emitter circuits," *IEEE MTT-S Int. Microwave Symp*, Dig., vol. 3, pp. 977-980, June 1999.

[10] V. Aparin and L.E. Larson, “Linearization of monolithic LNAs using low-frequency low-impedance input termination,” *European Solid-State Circuits Conference*, Sep. 2003, pp.137 – 140.

[11] K. L. Fong, “High-frequency analysis of linearity improvement technique of common-emitter trans-conductance stage using a low-frequency trap network,” *IEEE J. Solid-State Circuits*, vol. 35, no. 8, pp. 1249-1252, Aug. 2000.

[12] T. W. Kim, “A common-gate amplifier with transconductance nonlinearity cancellation and its high-frequency analysis using the Volterra series,” *IEEE Trans. Microw. Theory Tech.*, vol. 57, no. 6, pp. 1461–1469, June 2009.

[13] J. S. Fairbanks and Larson, L. E., “Analysis of optimized input and output harmonic termination on the linearity of 5 GHz CMOS radio frequency amplifiers,” *Radio and Wireless Conference*, Aug. 2003, pp. 293 – 296.

[14] T. W. Kim, B. Kim, and K. Lee, “Highly linear receiver front-end adopting MOSFET transconductance linearization by multiple gated transistors,” *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 223–229, Jan. 2004.

[15] X. Fan, H. Zhang, and E. Sánchez-Sinencio, “A noise reduction and linearity improvement technique for a differential cascode LNA,” *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 588-599, March 2008.

[16] V. Aparin, G. Brown, and L. E. Larson, "Linearization of CMOS LNAs via optimum gate biasing," in *IEEE Int. Circuits Syst. Symp.*, Vancouver, BC, Canada, vol. IV, pp. 748–751, May 2004.

[17] Y. Ding and R. Harjani, "A +18 dBm IIP3 LNA in 0.35μm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2001, pp. 162–163.

[18] E. Keehr, and A. Hajimiri, "Equalization of IM3 products in wideband direct-conversion receivers," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 204–205.

[19] Y. S. Youn, J. H. Chang, K. J. Koh, Y. J. Lee, and H. K. Yu, "A 2 GHz 16 dBm IIP3 low noise amplifier in 0.25 μm CMOS technology," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2003, pp. 452–453.

[20] C. Xin and E. Sánchez-Sinencio, "A linearization technique for RF low noise amplifier," in *IEEE Int. Circuits Syst. Symp.*, Vancouver, BC, Canada, vol. IV, pp. 313–316, May 2004.

[21] H. M. Geddada, J. W. Park and J. Silva-Martinez, "Robust derivative superposition method for linearizing broadband LNAs," *IEE Electronics Letters*, vol. 45 no. 9, pp.435-436, April 2009.

[22] D. Im, I. Nam, H. Kim, and K. Lee, "A wideband CMOS low noise amplifier employing noise and IM2 distortion cancellation for a digital TV tuner," *IEEE J. Solid-State Circuits*, vol. 44, No. 3, pp. 686-698, March 2009.

[23] T. W. Kim, and B. Kim, "A 13-dB IIP3 improved low-power CMOS RF programmable gain amplifier using differential circuit transconductanc linearization

for various terrestrial mobile D-TV applications," *IEEE J. Solid-State Circuits*, vol. 41, no. 4, pp. 945-953, April 2006.

[24] V. Aparin and L. E. Larson, "Modified derivative superposition method for linearizing FET low-noise amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 53, no. 2, pp. 571–581, Feb. 2005.

[25] S. Ganesan, E. Sánchez-Sinencio, and J. Silva-Martinez, "A highly linear low noise amplifier," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 12, pp. 4079-4085, Dec. 2006.

[26] S. Lou and H. C. Luong, "A linearization technique for RF receiver front-end using second-order-intermodulation injection," *IEEE J. Solid-State Circuits*, vol. 43, no. 11, pp. 2404-2412, Nov. 2008.

[27] F. Bruccoleri, E. A. M. Klumperink, and B. Nauta, "Wide-band CMOS low-noise amplifier exploiting thermal noise canceling," *IEEE J. Solid-State Circuits*, vol. 39, no. 2, pp. 275–282, Feb. 2004.

[28] J. Jussila*,* and P. Sivonen*,* "A 1.2-V highly linear balanced noise-cancelling LNA in 0.13-µm CMOS," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 579-587, Mar. 2008.

[29] W.Chen*,* G.Liu*,* B.Zdravko, and A.M. Niknejad, "A highly linear broadband CMOS LNA employing noise and distortion cancellation," *IEEE J. Solid-State Circuits*, vol. 43, no. 5, pp. 1164-1176, May 2008.

[30] S. C. Blaakmeer*,* E. A. M. Klumperink, D. M. W. Leenaerts*,* and B. Nauta*,* "Wideband balun-LNA with simultaneous output balancing, noise-canceling and

distortion-canceling," *IEEE J. Solid-State Circuits*, vol. 43, no. 6, pp. 1341-1350, June 2008.

[31] N. Kim, V. Aparin, K. Barnett, and C. Persico, "A cellular-band CDMA 0.25µm CMOS LNA linearized using active post-distortion," *IEEE J. Solid-State Circuits*, vol. 41, no. 7, pp. 1530–1534, Jul. 2006.

[32] T.-S. Kim and B.-S. Kim, "Post-linearization of cascode CMOS LNA using folded PMOS IMD sinker," *IEEE Microwave & Wireless Comp. Lett.*, vol. 16, no. 4, pp. 182-184, Apr. 2006.

[33] H. Zhang, X. Fan, and E. Sánchez-Sinencio, "A low-power, linearized, ultra-wideband LNA design technique," *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 320-330, Feb. 2009.

[34] D. Kaczman, M. Shah, M. Alam, M. Rachedine, D. Cashen, L. Han, and A. Raghavan, "A single-chip 10-band WCDMA/HSDPA 4-band GSM/EDGE SAW-less CMOS receiver with DigRF 3G interface and +90 dBm IIP2," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 718-739, March 2009.

[35] M. Hotti, J. Ryynanen, K. Kivekas, and K. Halonen, "An IIP2 calibration technique for direct conversion receivers," *in IEEE Int. Circuits Syst. Symp.,* Vancouver, BC, Canada, vol. IV, pp. 257–260, May 2004.

[36] W. Kim, S. Yang, Y. Moon, J. Yu, H. Shin, W. Choo, and B. Park, "IP2 calibrator using common mode feedback circuitry," *European Solid-State Circuits Conference*, Sep. 2005, pp.231 – 234.

[37] H. Darabi, H. Kim, J. Chiu, B. Ibrahim, and L. Serrano, "An IP2 improvement technique for zero-IF down-converters," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2006, pp. 1860-1869.

[38] W. Bosch and G. Gatti, "Measurement and simulation of memory effects in predistortion linearizers," *IEEE Trans. Microwave Theory Tech.*, vol. 37, pp. 1885–1890, Dec. 1989.

[39] J. F. Sevic, K. L. Burguer, and M. B. Steer, "A novel envelope-termination load–pull methods for ACPR optimization of RF/microwave power amplifiers," in *IEEE MTT-S Int. Microwave Symp*. Dig., Baltimore, MD, 1998, pp. 601–605.

[40] N. Borges de Carvalho and J. C. Pedro, "Two-tone IMD asymmetry in microwave power amplifiers," in I*EEE MTT-S Int. Microwave Symp*. Dig., Boston, MA, 2000, pp. 445–448.

[41] N. Carvalho, and J. Pedro, "A comprehensive explanation of distortion sideband asymmetries," *IEEE Trans. Microw. Theory Tech.*, vol. 50, no. 9, pp. 2090-2101, Sep. 2002.

[42] N. Cowley, and R. Hanrahan. (2005, Nov. 10). ATSC compliance and tuner design implications. *Video/Imaging Design Line* [Online]. Available: http://www.videsignline.com/showArticle.jhtml?articleID=173601582

[43] S. S. Taylor, and J. S. Duster, "High-linearity low noise amplifier and method," U. S. Patent 0 278 220, Nov. 13, 2008.

[44] S. Kang, B. Choi, and B. Kim, "Linearity analysis of CMOS for RF application," *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 3, pp. 972-977, Mar. 2003.

[45] R. A. Baki, T. K. K. Tsang, and M. N. El-Gamal, "Distortion in RF CMOS short-channel low-noise amplifiers," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 1, pp. 46-56, Jan. 2006.

[46] C. H. Choi, Z. Yu, and R. W. Dutton, "Impact of poly-gate depletion on MOS RF linearity," *IEEE Electron Device Lett.*, vol. 24, no. 5, pp. 330–332, May 2003.

[47] R. van Langevelde, L. F. Tiemeijer, R. J. Havens, M. J. Knitel, R. F. M. Ores, P. H.Woerlee, and D. B. M. Klaassen, "RF-distortion in deep submicron CMOS technologies," in *IEEE IEDM Tech. Dig.*, Dec. 2000, pp.807–810.

[48] T. Lee, and Y. Cheng, "High-frequency characterization and modeling of distortion behavior of MOSFETs for RF IC design," *IEEE J. Solid-State Circuits*, vol. 39, no.9, pp. 1407-1414, Sep. 2004.

[49] A. Annema, B. Nauta, R. Langevelde, and H. Tuinhout, "Analog Circuits in Ultra-Deep-Submicron CMOS," *IEEE J. of Solid-State Circuits*, vol. 40, no.1, pp. 132-143, Jan. 2005.

[50] M.S. Islam, and M.M. Zaman, "A seven-parameter nonlinear I-V characteristics model for sub-lm range GaAs MESFETs," *Solid-State Electronics*, vol. 48, no.7, pp. 1111–1117, July 2004.

[51] D. K. Sheffer, and T. H. Lee, "A 1.5V, 1.5GHz CMOS low-noise amplifier," *IEEE J. of Solid-State Circuits*, vol. 32, pp. 745-759, May 1997.

[52] T. K. Nguyen, C. H. Kim, G. J. Ihm, M. S. Yang, and S. G. Lee, "CMOS low-noise amplifier design optimization techniques," *IEEE Transactions on Microwave Theory and Techniques*, vol. 52, no. 5, pp. 1433-1442, May 2004.

[53] L. Belostotski and J. W. Haslett, "Noise figure optimization of inductively degenerated CMOS LNAs with integrated gate inductors," *IEEE Transaction on Circuits and Systems I: Regular Papers*, vol. 53, no.7, pp. 1409-1422, July 2006.

[54] H. Darabi and A. A. Abidi, "A 4.5mW 900-MHz CMOS receiver for wireless paging," *IEEE J. of Solid-State Circuits*, vol. 35, no. 8, pp. 1085-1096, Aug 2000.

[55] W. Zhuo, S. H. K. Embabi, J. Pineda de Gyvez, and E. Sánchez-Sinencio, "Using capacitive cross-coupling technique in RF low-noise amplifiers and down-conversion mixer design," *Proc. Eur. Solid-State Circuits Conf.*, pp.116-119, Sep. 2000.

[56] X. Li, S. Shekhar, and D. J. Allstot, "Gm-boosted common-gate LNA and differential colpitts VCO/QVCO in 0.18-µm CMOS," *IEEE J. of Solid-State Circuits*, vol. 40, no. 12, pp. 2609-2619, Dec 2005.

[57] W. Zhuo, X. Li, S. Shekhar, S. H. K. Embabi, J. Pineda de Gyvez, D. J. Allstot, and E. Sánchez-Sinencio, "A capacitor cross-coupled common-gate low noise amplifier," *IEEE Transaction on Circuits and Systems II: Express Briefs*, vol. 52, pp. 875-879, Dec 2005.

[58] T. H. Lee, H. Samavati, and H. R. Rategh, "5-GHz CMOS wireless LANs," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 1, pp. 268-280, Jan 2002.

[59] H. Samavati, H. R. Rategh, and T. H. Lee, "A 5-GHz CMOS wireless LAN receiver front end," *IEEE J. of Solid-State Circuits*, vol. 35, no. 5, pp. 765-772, May 2000.

[60] M. Zargari, M. Terrovitis, S. H.-M.Jen, B. J. Kaczynski, L. MeeLan, M. P. Mack, S.S. Mehta, S. Mendis, K. Onodera, H. Samavati, W. W.Si, K. Singh, A. Tabatabaei, D. Weber, D. K.Su, B. A.Wooley, "A single-chip dual-band tri-mode CMOS M. transceiver for IEEE 802.11a/b/g wireless LAN," *IEEE J. of Solid-State Circuits*, vol. 39, no. 12, pp. 2239-2249, Dec 2004.

[61] R. Fujimoto, K. Kojima, and S. Otaka, "A 7-GHz 1.8-dB NF CMOS low-noise amplifier," *IEEE J. of Solid-State Circuits*, vol. 37, no. 7, pp. 852-856, July 2002.

[62] W. Guo and D. Huang, "The noise and linearity optimization for a 1.9GHz CMOS low noise amplifier," *Proc. IEEE Asia-Pacific Conf. on ASIC*, pp. 253-257, Aug. 2002

[63] A. M. Niknejad, "ASITIC: Analysis and Simulation of Spiral Inductors and Transformers for ICs," Available at rfic.eecs.berkeley.edu/~niknejad/asitic.html [Aug.12, 2010]

[64] A. M. Niknejad, R. G. Meyer, "Analysis, design, and optimization of spiral inductors and transformers for Si RFIC's," *IEEE J. of Solid-State Circuits*, vol. 33, no. 10, pp. 1470-1481, Oct 1998.

[65] S. H. M. Lavasani, and S. Kiaei, "A new method to stabilize high frequency high gain CMOS LNA," in *IEEE Electronics, Circuits and Systems Conference*, Shadah, United Arab Emirates, pp. 982-985, Dec. 2003

[66] X. Li, T. Brogan, M. Esposito, B. Myers and K. O. Kenneth, "A comparison of CMOS and SiGe LNA's and mixers for wireless LAN application," *Proc. IEEE Custom Integrated Circuits Conf.*, 2001, pp. 531-534.

[67] V. Chandrasekhar, C. M. Hung, Y. C. Ho, and K. Mayaram, “A packaged 2.4GHz LNA in a 0.15μm CMOS process with 2kV HBM ESD protection,” *Proc. Eur. Solid-State Circuits Conf.*, pp. 347-350, Sept. 2002.

[68] J. Su, C. Meng, Y. Li, S. Tseng, and G. Huang, “2.4 GHz 0.35 um CMOS single-ended LNA and mixer with gain enhancement techniques,” in *Asia-Pacific on Microwave Conference Proceedings, APMC2005. Asia-Pacific Conference Proceedings*, Dec. 2005, pp. 1-4.

[69] C. Xin, and E. Sánchez-Sinencio, “A GSM LNA using mutual-coupled degeneration,” *IEEE Microwave and Wireless Components Letters,* vol. 15, no. 2 , pp. 68-70, Feb. 2005.

[70] A. Bevilacqua and A. M. Niknejad, “An ultrawideband CMOS low noise amplifier for 3.1-10.6GHz wireless receivers,” *IEEE J. Solid-State Circuits*, vol. 39, no.12, pp. 2259-2268, Dec. 2004.

[71] A. Ismail, and A. A. Abidi, “A 3-10GHz low noise amplifier with wideband LC-ladder matching network,” *IEEE J. Solid-State Circuits*, vol. 39, no.12, pp. 2269-2277, Dec 2004.

[72] X. Fan, E. Sánchez-Sinencio, and J. Silva-Martinez, “A 3GHz-10GHz common gate ultrawideband low noise amplifier,” *Proc. IEEE Midwest Symp. on Circuits and Systems,* pp. 631-634, Aug. 2005.

[73] K. Chen, J. Lu, B. Chen, and S. Liu, “An ultra-wide-band 0.4-10GHz LNA in 0.18μm CMOS,” *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 54, no. 3, pp. 217-221, March 2007.

[74] C. F. Liao and S. I. Liu, “A broadband noise-canceling MOS LNA for 3.1–10.6-GHz UWB receiver,” *IEEE J. Solid-State Circuits*, vol. 42, no. 2, pp. 329–339, Feb. 2007.

[75] S. Shekhar, J. S. Walling, and D. J. Allstot, “Bandwidth extension techniques for CMOS amplifiers,” *IEEE J. Solid-State Circuits*, vol. 41, no. 11, pp. 2424-2438, Nov. 2006.

[76] A. Valdes-Garcia, C. Mishra, F. Bahmani, J. Silva-Martinez, and E. Sánchez-Sinencio, “An 11-band 3.4 to 10.3GHz MB-OFDM UWB receiver in 0.25µm SiGe BiCMOS,” *IEEE J. Solid-State Circuits,* vol. 42, no. 4, pp. 935-948, Apr. 2007.

[77] B. Analui, and A. Hajimiri, “Bandwidth enhancement for transimpedance Amplifiers,” *IEEE J. Solid-State Circuits,* vol. 39, no. 8, pp. 1263-1270, Aug. 2004.

[78] M. T. Reiha*,* and J. R. Long*,* “A 1.2 V reactive-feedback 3.1–10.6 GHz low-noise amplifier in 0.13µm CMOS,” *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 1023-1033, May 2007.

[79] A. Amer, E. Hegazi, and H. Ragai, “A low-power wideband CMOS LNA for WiMax,” *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 54, no. 1, pp. 4-8, Jan. 2007.

[80] S. Chehrazi, A. Mirzaei, R. Bagheri, and A. A. Abidi, “A 6.5 GHz wideband CMOS low noise amplifier for multi-band use,” in *Proc. IEEE Custom Integrated Circuits Conf.*, Sep. 2005, pp. 801–804.

[81] C. S. Wang and C. K. Wang, “A 90 nm CMOS low noise amplifier using noise neutralizing for 3.1–10.6 GHz UWB system,” in *Proc. Eur. Solid-State Circuits Conf.*, Montreux, Switzerland, 2006, pp. 251–254.

[82] S. Lee, J. Bergervoet, K.S. Harish, D. Leenaerts, R. Roovers, R. van de Beek, and G. van der Weide, “A broadband receive chain in 65nm CMOS,” in *IEEE Int. Solid-State Circuits Conf.*, Feb. 2007, pp. 418–419.

[83] R. van de Beek, J. Bergervoet, H. Kundur, D. Leenaerts, and G. van der Weide, “A 0.6-to-10GHz receiver front-end in 45nm CMOS,” *IEEE Int. Solid-State Circuits Conf.*, Feb. 2008, pp. 128–129.

[84] J. M. de la Rosa, and M. Ismail, “Adaptive CMOS circuits for 4G wireless networks (Part II),” *IEEE European Conf. on Circuit Theory and Design,* Aug. 2007, pp.1-86.

[85] G. Gielen, E. Goris, and Y. Ke, “Reconfigurable A/D converters for flexible wireless transceivers in 4G radios,” in *Radio Design in Nanometer Technologies*, M. Ismail and D. Gonzalez, Springer, 2006, pp. 123-140.

[86] K. Gulati, “A low-power reconfigurable analog-to-digital converter,” PhD Dissertation, Massachusetts Institute of Technology, USA, 2001.

[87] J. Ryckaert, M. Verhelst, M. Badaroglu, S. D’Amico, V. De Heyn, et. al, “A CMOS ultra-wideband receiver for low data-rate communication,” *IEEE J. of Solid-State Circuits*, vol. 42, no. 11, pp. 2515-2527, Nov. 2007.

[88] J. Craninckx and G. Van der Plas, “A 65fJ/conversion-step 0-to-50MS/s 0-to-0.7mW 9b charge-sharing SAR ADC in 90nm digital CMOS,” *ISSCC Dig. Techn. Papers*, Feb. 2007, pp.246-247.

[89] N. Verma and A. P. Chandrakasan, “An ultra low energy 12-bit rate-resolution scalable SAR ADC for wireless sensor nodes,” *IEEE J. of Solid-State Circuits*, vol. 42, no. 6, pp.1196-1205, June 2007.

[90] T. N. Andersen, B. Hernes*,* A. Briskemyr, F. Telstø, J. Bjørnsen*,* T. E. Bonnerud, and Ø. Moldsvor, “A cost-efficient high-speed 12-bit pipeline ADC in 0.18-um digital CMOS,” *IEEE J. of Solid-State Circuits*, vol. 40, no. 7, pp.1506-1513, July 2005.

[91] I. Ahmed and D. A. Johns, “A 50-MS/s (35 mW) to 1-kS/s (15 µW) power scaleable 10-bit pipelined ADC using rapid power-on Opamps and minimal bias current variation,” *IEEE J. of Solid-State Circuits*, vol. 40, no. 12, pp.2446-2455, Dec. 2005.

[92] K. Iizuka, H. Matsui, M. Ueda, and M. Daito, “A 14bit digitally self-calibrated pipelined ADC with adaptive bias optimization for arbitrary speeds up to 40MS/s,” *IEEE J. of Solid-State Circuits*, vol. 41, no. 4, pp. 883-890, Apr. 2006.

[93] G. Geelen, E. Paulus, D. Simanjuntak, H. Pastoor and R. Verlinden, “A 90nm CMOS 1.2V 10b power and speed programmable pipelined ADC with 0.5pJ/conversion-step,” *ISSCC Dig. Techn. Papers*, pp. 214-215, 2006.

[94] C. Hsu, C. Huang, Y. Lin, C. Lee, Z. Soe, T. Aytur, and R. Yan, “A 7b 1.1GS/s reconfigurable time-interleaved ADC in 90nm CMOS,” in *Proc. of the Symposium on VLSI Circuits*, June 2007, pp.66-67.

[95] B. Xia, A. Valdes-Garcia, and E. Sánchez-Sinencio, "A 10-bit 44-MS/s 20mW configurable time-interleaved pipeline ADC for a dual-mode 802.11b/bluetooth receiver,” *IEEE J. of Solid-State Circuits*, vol. 41, no. 3, pp.530-539, March 2005.

[96] M. Anderson, K. Norling, A. Dreyfert, and J. Yuan, “A reconfigurable pipelined ADC in 0.18 um CMOS,” in *Proc. of the Symposium on VLSI Circuits*, June 2005, pp.326-329.

[97] T. Christen, T. Burger, and Q. Huang, “A 0.13um CMOS EDGE/UMTS/WLAN tri-mode ΔΣ ADC with -92dB THD,” *ISSCC Dig. Techn. Papers*, 2007, pp.240-241.

[98] S. Ouzounov, R. van Veldhoven, C. Bastiaansen, K. Vongehr, R. van Wegberg, et. Al, “A 1.2V 121-Mode CT ΔΣ modulator for wireless receivers in 90nm CMOS,” *ISSCC Dig. Techn. Papers*, 2007, pp.242-243.

[99] B. Putter, “A 5th-order CT/DT multi-mode ΔΣ modulator,” *ISSCC Dig. Techn. Papers*, 2007, pp. 244-245.

[100] P. Malla, H. Lakdawala, K. Kornegay, and K. Soumyanath, “A 28mW spectrum-sensing reconfigurable 20MHz 72dB-SNR DT ΔΣ ADC for 802.11n/WiMAX receivers,” *ISSCC Dig. Techn. Papers*, 2008, pp. 496-497.

[101] L. Bos, and V. U. Brussel, “A multirate 3.4-to-6.8mW 85-to-66dB DR GSM/Bluetooth/UMTS cascade DT ΔΣ modulator in 90nm digital CMOS,” *ISSCC Dig. Techn. Papers*, 2009, pp.176-177.

[102] H. Zhang, Q. Li and E. Sánchez-Sinencio, “Minimum current/area implementation of a cyclic ADC,” *IEE Electronics Letters*, vol. 45, no. 7, pp. 351-352, March 2009.

[103] B. Murmann and B. E. Boser, “A 12-bit 75-MS/s pipelined ADC using open-loop residue amplification,” *IEEE J. Solid-State Circuits,* vol. 38, no. 12, pp. 2040-2050, Dec. 2003.

[104] R. H. M. van Veldhoven, R. Rutten, L. J. Breems, “An inverter-based hybrid ΣΔ modulator,” *ISSCC Dig. Techn. Papers*, 2008, pp. 493-494.

[105] Y. Chae, and G. Han, “Low voltage, low power, inverter-based switched-capacitor delta-sigma modulator,” *IEEE J. Solid-State Circuits,* vol. 44, no. 2, pp. 458-472, Feb. 2009.

[106] L. Brooks and H. Lee, “A zero-crossing-based 8-bit 200 MS/s pipelined ADC,” *IEEE J. Solid-State Circuits*, vol. 42, no. 12, pp. 2677-2687, Dec. 2007.

[107] L. Brooks and H. Lee, “A 12b 50MS/s fully differential zero-crossing-based ADC without CMFB,” *ISSCC Dig. Techn. Papers*, Feb. 2009, pp. 166-167.

[108] J. Hu, N. Dolev, and B. Murmann, “A 9.4-bit, 50-MS/s, 1.44-mW pipelined ADC using dynamic residue amplification,” *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1057-1066, Apr. 2009.

[109] I. Ahmed, J. Mulder, and D. A. Johns, “A 50MS/s 9.9mW pipelined ADC with 58dB SNDR in 0.18μm CMOS using capacitive charge-pumps,” *ISSCC Dig. Techn. Papers*, Feb. 2009, pp. 164-165.

[110] E. Iroaga and B. Murmann, “A 12-Bit 75-MS/s pipelined ADC using incomplete sampling,” *IEEE J. Solid-State Circuits*, vol. 42, no. 4, pp. 748-756, Apr. 2007.

[111] N. Verma and A. P. Chandrakasan, “An ultra low energy 12-bit rate-resolution scalable SAR ADC for wireless sensor nodes,” *IEEE J. Solid-State Circuits*, vol. 42, no. 6, pp.1196-1205, June 2007.

[112] V. Giannini, P. Nuzzo, V. Chironi, A. Baschirotto, G. Plas, and J. Craninckx, “An 820μW 9b 40MS/s noise-tolerant dynamic-SAR ADC in 90nm digital CMOS,” *ISSCC Dig. Techn. Papers*, 2008, pp. 238-239.

[113] M. Elzakker, E. Tuijl, P. Geraedts, D. Schinkel, E. Klumperink, and B. Nauta, “A 1.9μW 4.4fJ/conversion-step 10b 1MS/s charge-redistribution ADC,” *ISSCC Dig. Techn. Papers*, 2008, pp. 244-245.

[114] A. Agnes, E. Bonizzoni, P. Malcovati, and F. Maloberti, “A 9.4-ENOB 1V 3.8μW 100kS/s SAR ADC with time-domain comparator,” *ISSCC Dig. Techn. Papers*, 2008, pp. 246-247.

[115] S. Naraghi, M. Courcy, and M. Flynn, “A 9b 14μW 0.06mm2 PPM ADC in 90nm digital CMOS,” *ISSCC Dig. Techn. Papers*, 2009, pp. 168-169.

[116] V. Dhanasekaran, M. Gambhir, M. M. Elsayed, E. Sanchez-Sinencio, J. Silva-Martinez, C. Mishra, L. Chen, and E. Pankratz, “A 20MHz signal bandwidth 68dB dynamic range continuous time ADC based on multi-bit time domain quantizer and feedback element,” *ISSCC Dig. Techn. Papers*, 2009, pp.174-175.

[117] M. Thompson, M. Werner, R. Egan, P. Coan, and P. Robl, “ Free running time to digital converter with 1 nanosecond resolution,” *IEEE Trans.on Nuclear Science,* vol. 35, no.1, pp. 184-186, Feb. 1988.

[118] P. Chen, C. Chen, J. Zheng, and Y. Shen, “ A PVT insensitive vernier-based time-to-digital converter with extended input range and high accuracy,” *IEEE Trans. on Nuclear Science,* vol. 54, no. 2, pp. 294-302, April 2007.

[119] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, and D. Schmitt-Landsiedel, “Variation tolerant high resolution and low latency time-to-digital converter,” in *European Solid State Circuits Conference, ESSCIRC Dig. Techn. Papers*, Sept. 2007, pp.194 – 197.

[120] V. Ramakrishnan and P. Balsara, “A wide-range, high-resolution, compact, CMOS time to digital converter,” *in International Conference on VLSI Design VLSID Dig. Tech. Papers,* Jan. 2006, pp. 6-11.

[121] T. Rahkonen and J. Kostamovaara, “The use of stabilized CMOS delay lines for the digitization of short time intervals,” *IEEE J. of Solid-State Circuits*, vol. 28, no. 8, pp.887–894, Aug. 1993.

[122] P. Chen, L. Shen-Luan, and W. Jingshown “A CMOS pulse-shrinking delay element for time interval measurement,” *IEEE Transactions on Circuits and Systems II.* vol. 47, no. 9, pp. 954–958, Sept. 2000.

[123] S. H. Lewis, H. S. Fetterman, G. F. Gross, Jr., R. Ramachandran, and T.R. Viswanathan, “A 10-b 20MS/s analog-to-digital converter,” *IEEE J. Solid-State Circuits*, vol. 27, no. 3, pp. 351-358, 1992.

[124] K. Nagaraj, H. S. Fetterman, J. Anidjar, S. H. Lewis, and R. G. Renninger, “A 250-mW, 8-b, 52MS/s paralell-pipelined A/D converter with reduced number of amplifiers,” *IEEE J. Solid-State Circuits*, vol. 32, no. 3, pp. 312-320, 1997.

[125] A. Valdes-Garcia*,* F. Abdel-Latif Hussien*,* J. Silva-Martínez*,* and E. Sánchez-Sinencio*,* “An integrated frequency response characterization system with a digital

interface for analog testing, " *IEEE J. Solid-State Circuits*, vol. 41, no. 10, pp. 2301-2313, 2006.

[126] L. Picolli, F. Maloberti, A. Rossini, F. Borghetti, P. Malcovati, and A. Baschirotto, "A 10-Bit pipeline A/D converter without timing signals," in *Proc. of* the *IEEE Int. Symp. on Circuits and Systems*, vol.1, May 2006, pp.5355-5358.

[127] A. Tasic, W. A. Serdijn, and J. R. Long, "Adaptive multi-standard circuits and systems for wireless communications," *IEEE Circuits and Systems Magazine*, vol. 6, no. 1, pp. 29-37, 2006.

[128] H. Zhang, M. M. Elsayed, and E. Sánchez-Sinencio, "New applications and technology scaling driving next generation A/D converters," in *IEEE European Conference on Circuit Theory and Design*, 2009, pp.109-112.

[129] I. Ahmed, and D. A. Johns, "A high bandwidth power scalable sub-sampling10-bit pipelined ADC with embedded sample and hold," *IEEE J. Solid-State Circuits*, vol. 43, no. 7, pp. 1638-1647, July 2008.

[130] J. Ryckaert, M. Verhelst, M. Badaroglu, S. D'Amico, V. De Heyn, et.al, "A CMOS ultra-wideband receiver for low data-rate communication," *IEEE J. Solid-State Circuits*, vol.42, no.11, pp. 2515-2527, Nov. 2007.

[131] W. Audoglio, E. Zuffetti, G. Cesura, and R. Castello, "A 6-10 bits reconfigurable 20MS/s digitally enhanced pipelined ADC for multi-standard wireless terminals," in *Proc. European Solid-State Circuits Conf.*, 2006, pp.496-499.

[132] G. Taylor and I. Galton, "A mostly digital variable-rate continuous-time ADC ΔΣ modulator," in *IEEE ISSCC Dig. Tech. Papers*, 2010, pp. 298-299.

[133] R. H. Walden, “Analog-to-digital converter survey and analysis,” *IEEE J. Selected Areas in Communications*, vol. 17, no. 4, pp.539-550, Apr. 1999.

[134] R. S. Assaad and J. Silva-Martinez, “The recycling folded cscode: A general enhancement of the folded cascode amplifier,” *IEEE J. Solid-State Circuits*, vol. 44, no. 9, pp. 2535-2542, Sep. 2009.

[135] D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl, and B. Nauta, “A double-tail latch-type voltage sense amplifier with 18ps setup+hold time,” in *IEEE ISSCC Dig. Tech. Papers*, 2007, pp. 314-315.

[136] B. Murmann, EE315B VLSI data conversion circuits lecture notes, [Online]. Available at: https://ccnet.stanford.edu/cgibin/course.cgi?cc=ee315b&action=handout_download&handout_id=ID125185307612754 [Aug.12, 2010]

APPENDIX A

TAYLOR COEFFICIENTS FOR NEGATIVE FEEDBACK SYSTEMS

A weakly nonlinear open-loop amplifier A with input X_e and output Y is modeled by:

$$Y = g_1 X_e + g_2 X_e^2 + g_3 X_e^3 \tag{A.1}$$

the 3rd-order closed-loop power series for Y_c is:

$$Y_c = b_1 X + b_2 X^2 + b_3 X^3 \tag{A.2}$$

To see how the negative feedback improves linearity, we should obtain the relation between b_i and g_i (i=1~3). Substituting $X_e = X - X_f = X - \beta Y$ into (A.1) yields:

$$\begin{aligned} Y &= g_1\left(X-\beta Y\right)+g_2\left(X-\beta Y\right)^2+g_3\left(X-\beta Y\right)^3 \\ &=\left(g_1-2g_2\beta Y+3g_3\beta^2Y^2\right)X+\left(g_2-3g_3\beta Y\right)X^2+g_3X^3-g_1\beta Y+g_2\beta^2Y^2-g_3\beta^3Y^3 \end{aligned} \tag{A.3}$$

By substituting (A.2) into (A.3) and neglecting 4th and higher order terms of X, we have:

$$\begin{aligned} &Y_c = Y \cong \left(g_1-g_1b_1\beta\right)X+\left(g_2-2g_2b_1\beta+g_2b_1^2\beta^2-g_1b_2\beta\right)X^2 \\ &+\left(g_3-2g_2b_2\beta+3g_3b_1^2\beta^2-g_1b_3\beta+2g_2b_1b_2\beta^2-g_3b_1^3\beta^3\right)X^3 \end{aligned} \tag{A.4}$$

We can equate the coefficients of X, X^2, and X^3 in (A.2) and (A.4) and solve the equations to obtain the closed loop coefficients as functions of the open loop coefficients:

$$b_1 = \frac{g_1}{1+T_0}, \quad b_2 = \frac{g_2}{\left(1+T_0\right)^3}, \quad b_3 = \frac{1}{(1+T_0)^4}\left(g_3 - \frac{2g_2^{\,2}}{g_1}\frac{T_0}{1+T_0}\right) \tag{A.5}$$

where $T_0 = g_1\beta$ is the linear open-loop gain.

APPENDIX B

VOLTERRA SERIES: INTRODUCTION & APPLICATIONS

B.1 Volterra Series: History

In 1887, Vito Volterra, the Italian mathematician and physicist, introduced " Volterra series" to model the nonlinear behavior; in 1942, Norbert Wiener, the American mathematician, applied Volterra series to analyze the nonlinear circuit; in 1957, J.F. Barrett systematically applied Volterra series to nonlinear systems, and later on D.A. George used the multidimensional Laplace transformation to study Volterra operators. Nowadays, Volterra series has been extensively used to calculated small, but nevertheless troublesome, distortion terms in transistor amplifiers and systems.

Why do we need Volterra series? Let's first introduce the concept of "Memory effect". In a system with memory effect, the output not only depends on the current input, but also on the previous inputs. Energy storage elements, e.g. capacitors and inductors, introduce memory effects. At low frequencies, there's enough time for charging/discharging before taking the output; however, at high frequencies, the output always contain a portion of the previous input due to insufficient discharging. Therefore, it is important to include memory effects for an accurate distortion analysis at high frequencies. However, Taylor series cannot capture memory effects, resulting in discrepancy in distortion analysis; on the other hand, Volterra series can predict more accurately these high-frequency-low-distortion terms for the "weakly nonlinear" time-invariant system with memory effects. Here the "weakly nonlinear" assumption means the input excitation is small, and polynomials can be used to model the nonlinearities. Note that Volterra series[4] may diverge thus become invalid for strongly nonlinear systems.

B.2 Volterra Series: Basics

A linear system without memory can be modeled as:

$$y(t) = h \cdot x(t) \tag{B.1}$$

where h is the linear gain, and output y at instant t only depends on input x at that time instant. A linear, discrete, causal, and time-invariant system with memory can be modeled as summing all the effects of past inputs with proper "weights":

$$y(n) = \sum_{i=0}^{n} h(\tau_i) \cdot x(n - \tau_i) \tag{B.2}$$

where n is the time index and h(τ) is the impulse response. For continuous time system, the convolution sum becomes a convolution integral:

$$y(t) = \int_0^t h(\tau) x(t - \tau) d\tau \tag{B.3}$$

For systems with 2nd-order nonlinearity, a memory-less system can be modeled as:

$$y_2(t) = h_2 \cdot x^2(t) \tag{B.4}$$

A system with memory can again be modeled as a weighted double sum with the 2nd-order impulse response as proper "weights":

$$y(n) = \sum_{j=0}^{n} \sum_{i=0}^{n} h_2\left(p_i, p_j\right) x(n - p_i) \cdot x(n - p_j) \tag{B.5}$$

In continuous time domain:

$$y_2(t) = \int\int_0^t h_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 \tag{B.6}$$

Now, we can generalize the expression for the nth-order nonlinear system. For memory-less systems represented using Taylor series:

$$y(t) = h_1 \cdot x(t) + h_2 \cdot x^2(t) + ... + h_n \cdot x^n(t) \tag{B.7}$$

System with memory represented with Volterra series is a sum of multidimensional convolution integrals:

$$\begin{aligned} y(t) &= H_1\left[x(t)\right] + H_2\left[x(t)\right] + H_3\left[x(t)\right] + ... + H_n\left[x(t)\right] \\ &= \int_0^t h_1(\tau_1) x(t-\tau_1) d\tau_1 + \int\int_0^t h_2(\tau_1, \tau_2) x(t-\tau_1) x(t-\tau_2) d\tau_1 d\tau_2 \\ &+ \iint\int_0^t h_3(\tau_1, \tau_2, \tau_3) x(t-\tau_1) x(t-\tau_2) x(t-\tau_3) d\tau_1 d\tau_2 d\tau_3 \\ &+ \cdots + \int ... \int_0^t h_n(\tau_1, \tau_2 ... \tau_n) x(t-\tau_1) x(t-\tau_2) ... x(t-\tau_n) d\tau_1 d\tau_2 ... d\tau_n \end{aligned} \tag{B.8}$$

where $H_n[x(t)] = \int ... \int h_n(\tau_1, \tau_2 ... \tau_n) x(t-\tau_1) x(t-\tau_2) ... x(t-\tau_n) d\tau_1 d\tau_2 ... d\tau_n$ is the nth-order Volterra operator, and $h_n(\tau_1, \cdots, \tau_n)$, the nth-order impulse response of the system, is the nth-order "Volterra kernel".

The above discussion is in time domain, to calculate the distortion such as HD$_{2,3}$ and IM$_{2,3}$, frequency domain Volterra kernels are needed, and the n-dimensional fourier transform can be used to obtain the nth-order frequency domain Volterra kernel H$_n$ from the time domain Volterr kernel:

$$\begin{aligned} H_n(\omega_1, \cdots, \omega_n) &= F\{h_n(\tau_1, \cdots, \tau_n)\} \\ &= \int \cdots \int h_n(\tau_1, \cdots, \tau_n) e^{-j\omega_1\tau_1} \cdots e^{-j\omega_n\tau_n} d\tau_1 ... d\tau_n \end{aligned} \tag{B.9}$$

For an input with m frequency components:

$$x = A(\cos\omega_1 t + \cos\omega_2 t + ... + \cos\omega_m t) \tag{B.10}$$

The output of the n^{th}-order nonlinear system can be denoted as:

$$y = H_1(j\omega_{p1}) \circ x + H_2(j\omega_{p1}, j\omega_{p2}) \circ x^2 + ... + H_n(j\omega_{p1}, j\omega_{p2}, ... j\omega_{pn}) \circ x^n \qquad \text{(B.11)}$$

where ω_{p1}, ω_{p2}, … , ω_{pn} are frequency variables, which would be substituted by the actual input signal frequencies; “o” is the Volterra operator, which contains both the amplitude multiplication and phase shift, i.e. each frequency component in x^n is multiplied by $\left|H_n(j\omega_{p1}, j\omega_{p2}, ... j\omega_{pn})\right|$, and the phase is shifted by $\angle H_n(j\omega_{p1}, j\omega_{p2}, ... j\omega_{pn})$. These phase shifting effect models the high frequency effects, which have been ignored in a memoryless Taylor series. For example, an input with two frequency components: $X = A(\cos\omega_1 t + \cos\omega_2 t)$, then the frequency variables ω_{p1}, ω_{p2} should be substituted by $\pm\omega_1$, $\pm\omega_2$, then $H_2(j\omega_{p1}, j\omega_{p2}) \circ X^2$ contains the following terms:

$$\left|H_2(j\omega_1, j\omega_1)\right| X^2 \angle H_2(j\omega_1, j\omega_1) = \frac{1}{2}\left|H_2(j\omega_1, j\omega_1)\right| A^2 \cos\left[2\omega_1 t + \angle H_2(j2\omega_1)\right]$$

$$\left|H_2(j\omega_1, -j\omega_1)\right| X^2 \angle H_2(j\omega_1, -j\omega_1) = \frac{1}{2}\left|H_2(j\omega_1, -j\omega_1)\right| A^2$$

$$\left|H_2(j\omega_1, j\omega_2)\right| X^2 \angle H_2(j\omega_1, j\omega_2) = \left|H_2(j\omega_1, j\omega_2)\right| A^2 \cos\left\{(\omega_1 + \omega_2)t + \angle H_2\left[j(\omega_1 + \omega_2)\right]\right\}$$

$$\left|H_2(j\omega_1, -j\omega_2)\right| X^2 \angle H_2(j\omega_1, -j\omega_2) = \left|H_2(j\omega_1, -j\omega_2)\right| A^2 \cos\left\{(\omega_1 - \omega_2)t + \angle H_2\left[j(\omega_1 - \omega_2)\right]\right\}$$

$$\left|H_2(j\omega_2, j\omega_2)\right| X^2 \angle H_2(j\omega_2, j\omega_2) = \frac{1}{2}\left|H_2(j\omega_2, j\omega_2)\right| A^2 \cos\left[2\omega_2 t + \angle H_2(j2\omega_2)\right]$$

Table B.1 compares the definition of distortion terms in Volterra series and Taylor series.

Table B.1 Definition of distortion terms

	Volterra series	Taylor series
HD2	$\frac{1}{2}\frac{\left\|H_2(j\omega_1, j\omega_1)\right\|}{\left\|H_1(j\omega_1)\right\|}A$	$\frac{1}{2}\frac{a_2}{a_1}A$
HD3	$\frac{1}{4}\frac{\left\|H_3(j\omega_1, j\omega_1, j\omega_1)\right\|}{\left\|H_1(j\omega_1)\right\|}A^2$	$\frac{1}{4}\frac{a_3}{a_1}A^2$
IM3	$\frac{3}{4}\frac{\left\|H_3(j\omega_1, j\omega_1, -j\omega_2)\right\|}{\left\|H_1(j\omega_1)\right\|}A^2$	$\frac{3}{4}\frac{a_3}{a_1}A^2$

The procedure for Volterra series analysis can be summarized as three steps:

Step 1: Define an intermediate variable v_i in terms of the input signal x:

$$v_i = G_1 \circ x + G_2 \circ x^2 + G_3 \circ x^3 \tag{B.12}$$

where G_i (i = 1-3) is the Volterra kernel relating x and v_i.

Step 2: Use KCL and MOS device equations to express output signal y in terms of x and vi. Solve G_i.

Step 3: Define output y in terms of input x:

$$y = H_1 \circ x + H_2 \circ x^2 + H_3 \circ x^3 \tag{B.13}$$

where H_i (i = 1-3) is the Volterra kernel relating x and y. H_i becomes a function of G_i, which has been determined in step 1. Substitute (B.12) into the equations obtained in step 2 to solve H_i. Section B.3 - B.6 illustrate these steps by showing four examples.

B.3 Volterra Series Analysis of the Common-Gate LNA (CG-LNA)

A typical CG input stage and its small signal equivalent circuit is shown in Fig. B.1, Here v_{in} and v_2 are the input and output, respectively.

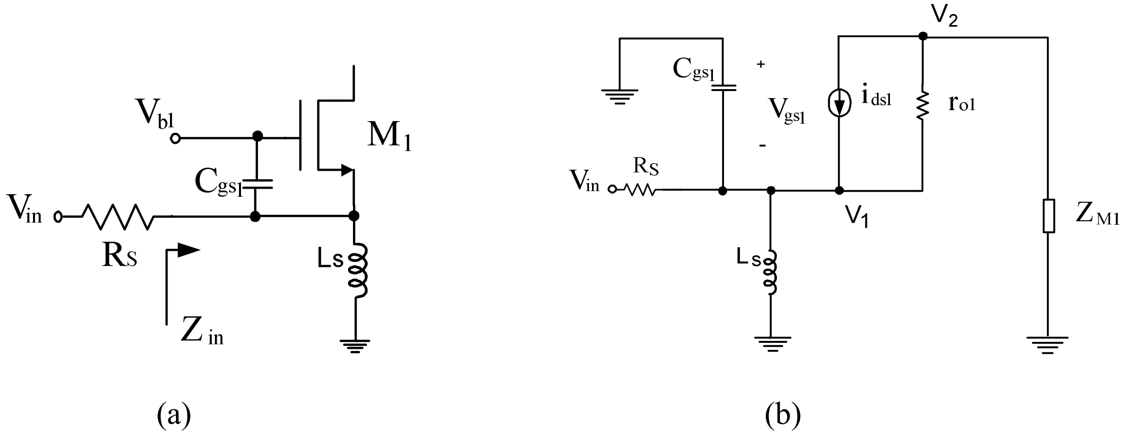


Fig. B.1. (a) Typical common gate LNA (b) small signal model

Step 1: define v_1, the voltage at the source node of M_1, as the intermediate variable, and express the relation between v_1 and v_{in} up to 3rd-order as:

$$v_1 = A_1(\omega) \circ v_{in} + A_2(\omega_1, \omega_2) \circ v_{in}^{\;2} + A_3(\omega_1, \omega_2, \omega_3) \circ v_{in}^{\;3} \tag{B.14}$$

Step 2: Write KCL equations for this circuit:

$$i_{ds1} = -g_{m1} v_1 + g_2 v_1^2 - g_3 v_1^3 \tag{B.15}$$

$$-\frac{v_2}{Z_{M1}} = \frac{v_1 - v_{in}}{R_s} + v_1\left(j\omega C_{gs1} + \frac{1}{j\omega L_s}\right) \tag{B.16}$$

$$i_{ds1} + \frac{v_2 - v_1}{r_{o1}} = \frac{v_1 - v_{in}}{R_s} + v_1\left(j\omega C_{gs1} + \frac{1}{j\omega L_s}\right) \tag{B.17}$$

Substituting (B.15) and (B.16) into (B.17) and cancel out i_{ds1} and v_2, we have:

$$r_{o1} \times \left(-g_{m1}v_1 + g_2 v_{\,1}^2 - g_3 v_1^3\right) = v_1 \times \left[1 + \left(r_{o1} + Z_{M1}\right)\left(\frac{1}{R_s} + j\omega C_{gs1} + \frac{1}{j\omega L_s}\right)\right] - \frac{1}{R_s}\left(r_{o1} + Z_{M1}\right)v_{in} \quad \text{(B.18)}$$

To obtain the expressions for the 1st-, 2nd-, and 3rd-order Volterra kernels $A_1(\omega)$, $A_2(\omega_1,\omega_2)$, and $A_3(\omega_1,\omega_2,\omega_3)$, we substitute (B.14) into (B.18) and cancel out v_1:

$$\begin{aligned}
&-g_{m1}r_{o1} \times \left[A_1(\omega)\circ v_{in} + A_2(\omega_1,\omega_2)\circ {v_{in}}^2 + A_3(\omega_1,\omega_2,\omega_3)\circ {v_{in}}^3\right] \\
&+g_2 r_{o1} \times \left[A_1(\omega)^2 \circ {v_{in}}^2 + 2\overline{A_1(\omega)A_2(\omega_1,\omega_2)}\circ {v_{in}}^3\right] - g_3 r_{o1} \times A_1(\omega)^3 \circ {v_{in}}^3 \\
&= \left[A_1(\omega)\circ v_{in} + A_2(\omega_1,\omega_2)\circ {v_{in}}^2 + A_3(\omega_1,\omega_2,\omega_3)\circ {v_{in}}^3\right] \\
&\times\left[1 + \left(r_{o1} + Z_{M1}\right)\left(\frac{1}{R_s} + j\omega C_{gs1} + \frac{1}{j\omega L_s}\right)\right] - \frac{1}{R_s}\left(r_{o1} + Z_{M1}\right)v_{in}
\end{aligned} \quad \text{(B.19)}$$

where $\overline{A_1(\omega_1)A_2(\omega_1,\omega_2)} = \frac{1}{3}\left[A_1(\omega_1)A_2(\omega_2,\omega_3) + A_1(\omega_2)A_2(\omega_1,\omega_3) + A_1(\omega_3)A_2(\omega_1,\omega_2)\right]$.

To get $A_1(\omega)$, we assume a single input tone, i.e. $v_{in} = e^{\omega t}$. By equating the coefficients of $e^{\omega t}$ of (B.19), we can get:

$$A_1(\omega) = \frac{r_{o1} + Z_{M1}}{H(\omega)} \quad \text{(B.20)}$$

where $H(\omega) = R_s + r_{o1} + Z_{M1} + g_{m1}r_{o1}R_s + R_s B(\omega)\left(r_{o1} + Z_{M1}\right)$, $B(\omega) = j\omega C_{gs1} + 1/j\omega L_s$

Apply the two tone input $v_{in} = e^{\omega_1 t} + e^{\omega_2 t}$ to (B.19) and equate its coefficients, we can get:

$$A_2(\omega_1,\omega_2) = \frac{g_2 \cdot a_1^2 \cdot R_s \cdot r_{o1}}{H(\omega_1 + \omega_2)} \quad \text{(B.21)}$$

Apply the three tone input $v_{in} = e^{\omega_1 t} + e^{\omega_2 t} + e^{\omega_3 t}$ to (B.19) and equate its coefficients:

$$A_3(\omega_1,\omega_2,\omega_3) = \frac{A_1^3 \cdot R_s \cdot r_{o1} \cdot \varepsilon\left(\Delta\omega, \omega_1 + \omega_2\right)}{H\left(\omega_1 + \omega_2 + \omega_3\right)} \quad \text{(B.22)}$$

where

$$\varepsilon(\Delta\omega,\omega_1+\omega_2)=g_3-g_{oB}(\Delta\omega,\omega_1+\omega_2),\ \text{ and }\ g_{oB}(\Delta\omega,\omega_1+\omega_2)=\frac{2}{3}g_2{}^2 r_{o1}R_s\left[1/H(\Delta\omega)+1/H(\omega_1+\omega_2)\right]$$

Step 3: express the relation between v_2 and v_{in} using Volterra seires:

$$v_2=C_1(\omega)\circ v_{in}+C_2(\omega_1,\omega_2)\circ v_{in}{}^2+C_3(\omega_1,\omega_2,\omega_3)\circ v_{in}{}^3 \tag{B.23}$$

Substituting (B.14) and (B.23) into (B.16), we have:

$$\begin{aligned}&C_1(\omega)\circ v_{in}+C_2(\omega_1,\omega_2)\circ v_{in}{}^2+C_3(\omega_1,\omega_2,\omega_3)\circ v_{in}{}^3\\&=-Z_{M1}\cdot\left\{\begin{aligned}&v_{in}\cdot\left[A_1(\omega)\left(\frac{1}{R_s}+j\omega C_{gs1}+\frac{1}{j\omega L_s}\right)-\frac{1}{R_s}\right]\\&+\left(\frac{1}{R_s}+j\omega C_{gs1}+\frac{1}{j\omega L_s}\right)\times\left[A_2(\omega_1,\omega_2)\circ v_{in}{}^2+A_3(\omega_1,\omega_2,\omega_3)\circ v_{in}{}^3\right]\end{aligned}\right\}\end{aligned} \tag{B.24}$$

By applying $v_{in} = e^{\omega t}$, $v_{in} = e^{\omega_1 t}+e^{\omega_2 t}$, and $v_{in} = e^{\omega_1 t}+e^{\omega_2 t}+e^{\omega_3 t}$ into (B.24) and equating their coefficients respectively, we can have:

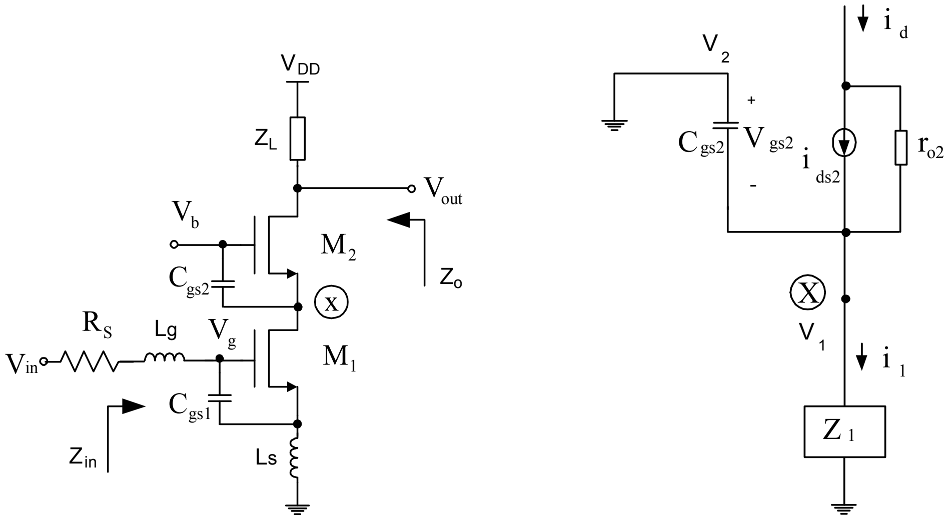
$$C_1(\omega)=-Z_{M1}\cdot A_1(\omega)\cdot\left(\frac{1}{R_s}+B(\omega)\right)+Z_{M1}\cdot\frac{1}{R_s} \tag{B.25}$$

$$C_2(\omega_1,\omega_2)=\frac{-Z_{M1}\cdot\left[1+R_s\cdot\left(j\omega C_{gs1}+\dfrac{1}{j\omega L_s}\right)\right]\cdot A_2(\omega_1,\omega_2)}{R_s} \tag{B.26}$$

$$C_3(\omega_1,\omega_2,\omega_3)=\frac{-Z_{M1}\cdot A_3(\omega_1,\omega_2,\omega_3)\cdot\left(1+R_sB(\omega)\right)}{R_s} \tag{B.27}$$

B.4. Volterra Series Analysis of the Common-Source LNA (CS-LNA) with Cascode

Fig. B.2 shows a typical CS-LNA with cascode stage and its small signal model.



(a) (b)

Fig. B.2. (a) Typical common source LNA with cascode (b) Small signal model

Applying KCL to each node of the model in Fig.B.2, we can get:

$$j\omega C_{gs2}(V_2 - V_1) + i_d = j\omega C_{gs2} V_{gs2} + i_d = i_1 \tag{B.28}$$

$$V_1 = i_1 \times Z_1 \tag{B.29}$$

where i_1 is the input, and i_d is the output. The relation between i_1 and i_d can be expressed up to 3rd-order using Volterra series as:

$$i_d = i_{ds2} = f(i_1) = c_1(\omega) \circ i_1 + c_2(\omega_1, \omega_2) \circ i_1^2 + c_3(\omega_1, \omega_2, \omega_3) \circ i_1^3 \tag{B.30}$$

Express the relation between the drain currents of M2 and the gate source voltage V_{gs2} up to 3rd-order:

$$f\left(V_{gs}\right) = i_{ds2} \approx g_m V_{gs2} + g_2 V_{gs2}{}^2 + g_3 V_{gs2}{}^3 \tag{B.31}$$

Express the relation between V_{gs2} and the input i_1 up to 3rd-order with Volterra series as:

$$V_{gs2} \approx a_1(\omega) \circ i_1 + a_2(\omega_1, \omega_2) \circ i_1^2 + a_3(\omega_1, \omega_2, \omega_3) \circ i_1^3 \tag{B.32}$$

where $a_1(\omega)$ is the 1st-order coefficient with one input frequency, $a_2(\omega_1, \omega_2)$ is the 2nd-order coefficient with two input frequencies and $a_3(\omega_1, \omega_2, \omega_3)$ is the 3rd-order coefficient with three input frequencies. They represent the mixed nonlinear effect for multiple input frequencies. $a_1(\omega)$, $a_2(\omega_1, \omega_2)$ and $a_3(\omega_1, \omega_2, \omega_3)$ can be obtained by solving (B.28)-(B.32) by equating the same order terms of i_1 at both sides of the equations.

Substituting (B.31) into (B.32), we can get

$$\begin{aligned} i_d \approx i_{ds2} \approx\ & g_m a_1(\omega) \circ i_1 + [g_m a_2(\omega_1, \omega_2) + g_2 a_1(\omega_1) a_1(\omega_2)] \circ i_1^2 \\ & + [g_m a_3(\omega_1, \omega_2, \omega_3) + 2 g_2 \overline{a_1(\omega_1) a_2(\omega_2, \omega_3)} + g_3 a_1(\omega_1) a_1(\omega_2) a_1(\omega_3)] \circ i_1^3 \end{aligned} \tag{B.33}$$

where

$$\overline{a_1(\omega_1) a_2(\omega_2, \omega_3)} = \frac{1}{3}[a_1(\omega_1) a_2(\omega_2, \omega_3) + a_1(\omega_2) a_2(\omega_1, \omega_3) + a_1(\omega_3) a_2(\omega_1, \omega_2)] \tag{B.34}$$

Substituting (B.32), (B.34) into (B.28), we can get

$$\begin{aligned} & j\omega C_{gs2} \left(a_1(s) \circ i_1 + a_2(s_1, s_2) \circ i_1^2 + a_3(s_1, s_2, s_3) \circ i_1^3 \right) \\ & + g_m a_1(s) \circ i_1 + \left(g_1 a_2(s_1, s_2) + g_2 a_1(s_1) a_1(s_2) \right) \circ i_1^2 \\ & + \left(g_m a_3(s_1, s_2, s_3) + 2 g_2 \overline{a_1(s_1) a_2(s_2, s_3)} + g_3 a_1(s_1) a_1(s_2) a_1(s_3) \right) \circ i_1^3 = i_1 \end{aligned} \tag{B.35}$$

For the harmonic input method, (B.35) needs to hold true for all the 1st-, 2nd-, and 3rd-order terms. With a single input tone, $i_1 = e^{\omega t}$, equating the coefficients of $e^{\omega t}$ of (B.35), we can get

$$a_1(\omega) = \frac{1}{g_m + j\omega C_{gs2}} \tag{B.36}$$

Applying the two tones input, $i_1 = e^{\omega_1 t} + e^{\omega_2 t}$, to (B.35) and equating the coefficients of $e^{(\omega_1+\omega_2)t}$, we can get

$$a_2(\omega_1,\omega_2) = -\frac{-g_2 a_1^2(\omega)}{g_m + \left(\omega_1 + \omega_2\right)C_{gs2}} \tag{B.37}$$

Applying the three tones input, $i_1 = e^{\omega_1 t} + e^{\omega_2 t} + e^{\omega_3 t}$, to (B.35) and equating the coefficients of $e^{(\omega_1+\omega_2+\omega_3)t}$, we can get

$$a_3(\omega_1,\omega_2,\omega_3) = -\frac{-2g_2\overline{a_1(\omega_1)a_2(\omega_1,\omega_2)}}{g_m + \left(\omega_1 + \omega_2 + \omega_3\right)C_{gs2}} + g_3 a_1{}^3(\omega) \tag{B.38}$$

Substituting (B.35)-(B.38) into (B.30) and (B.31), we can get

$$c_1(\omega) = g_m a_1(\omega) \tag{B.39}$$

$$c_2(\omega) = g_m a_2(\omega_1,\omega_2) + g_2 a_1(\omega_1)a_1(\omega_2) \tag{B.40}$$

$$c_3(\omega) = g_m a_3(\omega_1,\omega_2,\omega_3) + 2g_2\overline{a_1(\omega_1)a_2(\omega_2,\omega_3)} + g_3 a_1(\omega_1)a_1(\omega_2)a_1(\omega_3) \tag{B.41}$$

The A_{IIP3} of the cascode stage can be derived as

$$\mathrm{A}_{\mathrm{IIP3}}^2 = \frac{4}{3}\cdot\frac{1}{|\mathrm{H}(\omega)|\cdot|\mathrm{a}_1(\omega)|^3\cdot|\varepsilon(\Delta\omega,2\omega)|} \tag{B.42}$$

$$\mathrm{g}(\omega) = \mathrm{j}\omega\mathrm{C}_{\mathrm{gs2}} \tag{B.43}$$

$$\varepsilon(\Delta\omega,2\omega) = \mathrm{g}_3 - \mathrm{g}_{\mathrm{oB}} \tag{B.44}$$

$$\mathrm{g}_{\mathrm{oB}} = \frac{2}{3}\mathrm{g}_2^2\left[\frac{2}{\mathrm{g}_\mathrm{m} + \mathrm{g}(\Delta\omega)} + \frac{1}{\mathrm{g}_\mathrm{m} + \mathrm{g}(2\omega)}\right] \tag{B.45}$$

$$\mathrm{H}(\omega) = \frac{\mathrm{g}(\omega)}{\mathrm{g}_\mathrm{m}} \tag{B.46}$$

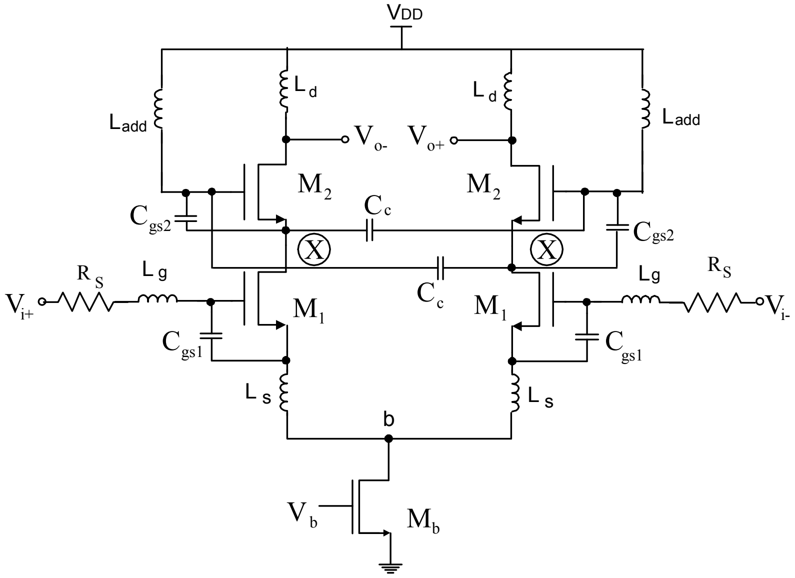


Fig. B.3. The proposed differential cascode CS-LNA.

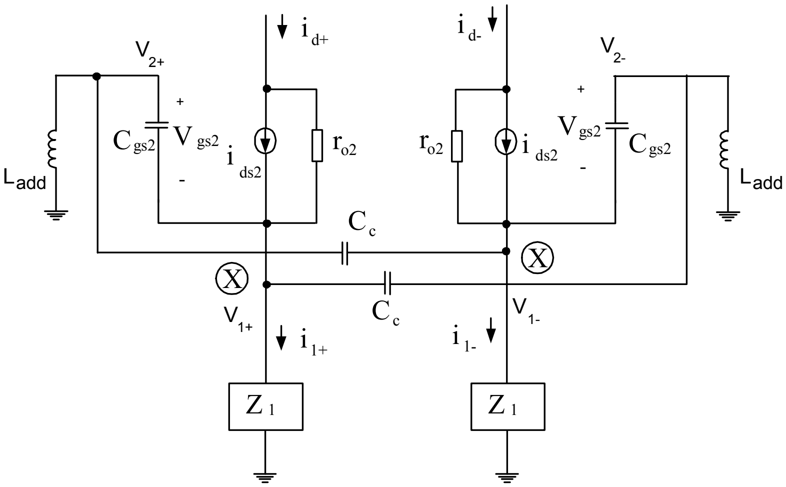


Fig. B.4. Analyzed cascode stage equivalent circuit.

Fig.B.3 and Fig. B.4 show the proposed differential cascode CS-LNA and the small-signal circuit for Volterra series analysis, respectively. Applying KCL to every node of the model in Fig. B.4:

$$j\omega C_{gs2}(V_{2+}-V_{1+})+i_{d+}+j\omega C_c(V_{2-}-V_{1+})=i_{1+} \tag{B.47}$$

$$j\omega C_c(V_{1-}-V_{2+})=j\omega C_{gs2}(V_{2+}-V_{1+})+j\omega L_{add}V_{2+} \tag{B.48}$$

$$\mathrm{V}_{1-}=-\mathrm{V}_{1+} \tag{B.49}$$

$$\mathrm{V}_{2-}=-\mathrm{V}_{2+} \tag{B.50}$$

$$\mathrm{i}_{1-}=-\mathrm{i}_{1+} \tag{B.51}$$

$$\mathrm{i}_{d-}=-\mathrm{i}_{d+} \tag{B.52}$$

For the cascode stage with the proposed technique, we can get

$$\mathrm{i}_d=\mathrm{i}_1-\mathrm{g}^{'}(\omega)\cdot \mathrm{V}_{gs2} \tag{B.53}$$

$$\mathrm{g}^{'}(\omega)=\frac{4\mathrm{j}\omega \mathrm{C}_{gs2}\cdot \mathrm{j}\omega \mathrm{C}_c+\frac{1}{\mathrm{j}\omega \mathrm{L}_{add}}(\mathrm{j}\omega \mathrm{C}_{gs2}+\mathrm{j}\omega \mathrm{C}_c)}{2\mathrm{j}\omega \mathrm{C}_c+\frac{1}{\mathrm{j}\omega \mathrm{L}_{add}}}+\omega \mathrm{C}_{sb2}+\omega \mathrm{C}_{gd1}+\omega \mathrm{C}_{db1} \tag{B.54}$$

Replacing (B.43) with (B.54), all the other results from (B.42)-(B.46) are still valid. For the proposed technique, if (B.54) equals to zero, the current generated by M_1 will all flow to the output without nonlinearity degradation. It helps to improve the LNA linearity.

For the typical CS-LNA with a cascode transistor, the nonlinearity degradation can be evaluated by (B.42). From DC simulation, calculate the gate source capacitance C_{gs2}, the 1^{st}-order transconductance gm, the 2^{nd}- and the 3^{rd}-order nonlinearity term g_2

and g_3. Calculate $g(\omega)$, g_{oB}, $\varepsilon(\Delta\omega,2\omega)$ and $H(\omega)$ using (B.43)-(B.46). Calculate the input 3rd-order intermodulation using (B.42).

B.5. Derivation of Volterra Kernels for Negative Feedback Systems

Model the weakly nonlinear amplifier A with input X_e and output Y in Volterra series up to 3rd-order as:

$$Y = g_1(\omega) \circ X_e + g_2(\omega_1,\omega_2) \circ X_e^2 + g_3(\omega_1,\omega_2,\omega_3) X_e^3 \tag{B.55}$$

the 3rd-order closed-loop Volterra series for Y_c is:

$$Y_c = b_1(\omega) \circ X + b_2(\omega_1,\omega_2) \circ X^2 + b_3(\omega_1,\omega_2,\omega_3) X^3 \tag{B.56}$$

Substituting $X_e = X - X_f = X - \beta(\omega) \circ Y$ into (B.55) yields:

$$\begin{aligned} Y =& \left[g_1(\omega) - 2g_2(\omega_1,\omega_2)\beta(\omega) \circ Y + 3g_3(\omega_1,\omega_2,\omega_3)\beta(\omega_1+\omega_2) \circ Y^2 \right] \circ X \\ &+ \left[g_2(\omega_1,\omega_2) - 3g_3(\omega_1,\omega_2,\omega_3)\beta(\omega) \circ Y \right] \circ X^2 + g_3(\omega_1,\omega_2,\omega_3) \circ X^3 \\ &- g_1(\omega)\beta(\omega) \circ Y + g_2(\omega_1,\omega_2)\beta(\omega_1+\omega_2) \circ Y^2 - g_3(\omega_1,\omega_2,\omega_3)\beta(\omega_1+\omega_2+\omega_3) \circ Y^3 \end{aligned} \tag{B.57}$$

Substitute (B.56) into (B.57) and neglecting 4th and higher order terms of X, we have:

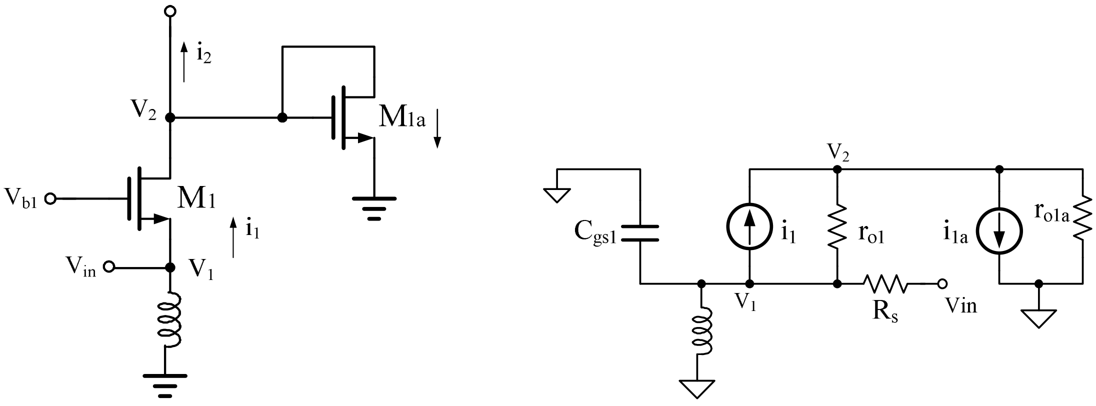
$$\begin{aligned} Y_c = Y \cong& \left[g_1(\omega) - g_1(\omega) b_1(\omega) \beta(\omega) \right] \circ X \\ &+ \begin{bmatrix} g_2(\omega_1,\omega_2) - 2g_2(\omega_1,\omega_2) b_1(\omega) \beta(\omega) \\ +g_2(\omega_1,\omega_2) b_1(\omega_1+\omega_2) \beta(\omega_1+\omega_2) - g_1(\omega) b_2(\omega_1,\omega_2) \beta(\omega) \end{bmatrix} \circ X^2 \\ &+ \begin{bmatrix} g_3(\omega_1,\omega_2,\omega_3) - 2g_2(\omega_1,\omega_2) b_2(\omega_1,\omega_2) \beta(\omega) \\ +3g_3(\omega_1,\omega_2,\omega_3) b_1(\omega_1+\omega_2) \beta(\omega_1+\omega_2) \\ -g_1(\omega) b_3(\omega_1,\omega_2,\omega_3) \beta(\omega) + 2g_2(\omega_1,\omega_2) b_1(\omega) b_2(\omega_1,\omega_2) \beta(\omega_1+\omega_2) \\ -g_3(\omega_1,\omega_2,\omega_3) b_1(\omega_1+\omega_2+\omega_3) \beta(\omega_1+\omega_2+\omega_3) \end{bmatrix} \circ X^3 \end{aligned} \tag{B.58}$$

To obtain the 3rd-order closed loop Volterra kernels, $b_3(\omega_1, \omega_2, \omega_3)$, as a function of the open loop Volterra kernels, we can apply a three-tone-input X = $e^{\omega_1 t} + e^{\omega_2 t} + e^{\omega_3 t}$ to the system. By equating the coefficients of $e^{(\omega_1+\omega_2+\omega_3)t}$, we have:

$$b_3(\omega_1,\omega_2,\omega_3) = \frac{1}{\big(1+T(\omega_1+\omega_2+\omega_3)\big)\big(1+T(\omega_1)\big)\big(1+T(\omega_2)\big)\big(1+T(\omega_3)\big)} \times$$
$$\left[g_3(\omega_1,\omega_2,\omega_3) - \frac{2g_2(\omega_1+\omega_2)}{3g_1(\omega_1)}\left(\frac{T(\omega_2+\omega_3)}{1+T(\omega_2+\omega_3)} + \frac{T(\omega_1+\omega_3)}{1+T(\omega_1+\omega_3)} + \frac{T(\omega_1+\omega_2)}{1+T(\omega_1+\omega_2)}\right)\right] \quad \text{(B.59)}$$

B.6. Volterra Series Analysis of the Proposed Linearized UWB CG-LNA

Fig. B.5 shows the circuit and small-signal model of the proposed linearized UWB CG-LNA for Volterra series analysis.



(a) (b)

Fig. B.5 (a) Proposed linearized UWB CG-LNA (b) small-signal equivalent circuit for linearity analysis

Express the relation between v_1 and v_{in}, v_2 and v_1, up to 3rd-order as:

$$v_1 = A_1(\omega) \circ v_{in} + A_2(\omega_1, \omega_2) \circ v_{in}{}^2 + A_3(\omega_1, \omega_2, \omega_3) \circ v_{in}{}^3 \tag{B.60}$$

$$v_2 = b_1(\omega) \circ v_1 + b_2(\omega_1, \omega_2) \circ v_1{}^2 + b_3(\omega_1, \omega_2, \omega_3) \circ v_1{}^3 \tag{B.61}$$

Write KCL equations for this circuit:

$$i_1 = g_{m1} v_1 + g_2 v_1^2 + g_3 v_1^3 \tag{B.62}$$

$$i_{1a} = g_m^{'} v_2 + g_2^{'} v_2^2 + g_3^{'} v_2^3 \tag{B.63}$$

$$i_2 = i_1 - i_{1a} \tag{B.64}$$

$$\frac{v_2 - v_1}{r_{o1}} - i_1 = \frac{v_1 - v_{in}}{R_s} + v_1 \left(j\omega C_{gs1} + \frac{1}{j\omega L_s} \right) \tag{B.65}$$

$$i_1 + \frac{v_2 - v_1}{r_{o1}} = i_{1a} + \frac{v_2}{r_{o1a}} \tag{B.66}$$

By solving equations (B.60)-(B.66), we can get:

$$\begin{aligned}
&\left(g_m - b_1 g_{ma}\right) \cdot A_1 v_{in} + \left[\left(g_m - b_1 g_{ma}\right) \cdot A_2 + \left(g_2 - b_1^2 g_{2a} - b_2 g_{ma}\right) \cdot A_1^2\right] \cdot v_{in}^2 \\
&+ \left[\left(g_m - b_1 g_{ma}\right) \cdot A_3 + \left(g_2 - b_1^2 g_{2a} - b_2 g_{ma}\right) \cdot 2A_1 A_2 + \left(g_3 - b_1^3 g_{3a} - b_3 g_{ma} - 2 g_{2a} b_1 b_2\right) \cdot A_1^3\right] \cdot v_{in}^3 \\
&= \left[\alpha A_1 - \gamma + (1+\beta) g_m A_1\right] \cdot v_{in} + \left[\alpha A_2 + (1+\beta)\left(g_m A_2 + g_2 A_1^2\right)\right] \cdot v_{in}^2 \\
&+ \left[\alpha A_3 + (1+\beta)\left(g_m A_3 + 2 g_2 A_1 A_2 + g_3 A_1^3\right)\right] \cdot v_{in}^3
\end{aligned} \tag{B.67}$$

where $\beta = \dfrac{r_{o1}}{r_{o1a}}$, $\gamma = \dfrac{1}{R_s}(1+\beta)$, $\alpha = \dfrac{1}{r_{o1a}} + B(1+\beta) + \gamma$, and $B(\omega) = j\omega C_{gs1} + \dfrac{1}{j\omega L_s}$

Apply a single input tone $v_{in} = e^{\omega t}$ and equate the coefficients of $e^{\omega t}$ in (B.67), we have:

$$A_1(\omega) = \frac{1}{R_s} \cdot \frac{1 + r_{o1}/r_{o1a}}{H(\omega)} \tag{B.68}$$

$$H(\omega) = \left(\frac{1}{R_s} + B(\omega)\right)\left(1 + \frac{r_{o1}}{r_{o1a}}\right) + \frac{1}{r_{o1a}} + \frac{r_{o1}}{r_{o1a}} g_m + b_1 g_{ma} \tag{B.69}$$

Apply a two-tone input $v_{in} = e^{\omega_1 t} + e^{\omega_2 t}$ and equate the coefficients of $e^{(\omega_1 + \omega_2)t}$ in (B.67), we can get:

$$A_2(\omega_1 + \omega_2) = \frac{-\left(\frac{r_{o1}}{r_{o1a}} g_2 + b_1^2 g_{2a} + b_2 g_{ma}\right) \cdot A_1^2}{H(\omega_1 + \omega_2)} \tag{B.70}$$

Apply a three-tone input $v_{in} = e^{\omega_1 t} + e^{\omega_2 t} + e^{\omega_3 t}$ and equate the coefficients of $e^{(\omega_1 + \omega_2 + \omega_3)t}$ in (B.67), we can get:

$$A_3(\omega_1, \omega_2, \omega_3) = -\frac{\left(\frac{r_{o1}}{r_{o1a}} g_3 + b_1^3 g_{3a} + b_3 g_{ma} + 2 b_1 b_2 g_{2a}\right) \cdot A_1^3 + 2\overline{A_1 A_2}\left(b_1^2 g_{2a} + b_2 g_{ma} + \frac{r_{o1}}{r_{o1a}} g_2\right)}{H(\omega_1 + \omega_2 + \omega_3)} \tag{B.71}$$

$$i_{2,3rd} = A_1(\omega)^3 \cdot \left[\left(g_3 - b_1^{\ 2} g_{3a}\right) - \frac{g_3^{'}\left(g_m - b_1 g_{ma}\right)}{H(\omega)} + \frac{g_{oB}^{'}\left(\Delta\omega, \omega_1 + \omega_2\right)\left(g_m - b_1 g_{ma}\right)}{H(\omega)} - \frac{2}{3}\frac{\left(g_2^{\ 2} - b_1^{\ 4} g_{2a}^{\ \ 2}\right)}{H(\omega_1 + \omega_2)}\right] \cdot v_{in}^{\ 3} \tag{B.72}$$

where $g_3^{'} = \frac{r_{o1}}{r_{o1a}} g_3 + b_1^{\ 3} g_{3a}$, $g_{oB}^{'}\left(\Delta\omega, \omega_1 + \omega_2\right) = \frac{2}{3}\left(\frac{r_{o1}}{r_{o1a}} g_2 + b_1^{\ 2} g_{2a}\right)^2 \left[\frac{1}{H(\Delta\omega)} + \frac{1}{H(\omega_1 + \omega_2)}\right]$

VITA

Heng Zhang received her B.S. degree in electrical engineering from Peking University, Beijing, China, in 2004, and her Ph.D. degree in electrical engineering from Texas A&M University, College Station, in 2010.

During the summer and fall of 2006, she was an Analog IC Design Engineer (Co-op) with Texas Instrument, Dallas, TX, where she designed a low power ADC for hard disk applications. In the summer 2007, she was with the RF and Analog Technologies Department, UMC, Sunnyvale, CA, where she researched digital calibration techniques for ADCs. In the summer 2008, she received a business management certificate in Mays Business School of Texas A&M University. Since August 2010, she has been with the Analog and Mixed Signal Group, Broadcom Corporation, Irvine, CA, working on high-speed transceivers for optical and backplane/cable applications. Her research interests include data converters and high-speed/RF circuits. Her website can be found at http://www.ece.tamu.edu/~hzhang

Ms. Zhang can be reached through the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128. Her email address is: hengzhangpku@hotmail.com