Built-in-Self-Test and Digital Self-Calibration for Radio Frequency Integrated Circuits

DISSEPTION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Sleiman Bou Sleiman, M.Sc.

Graduate Program in Electrical and Computer Engineering

The Ohio State University

2011

Dissertation Committee:

Prof. Mohammed Ismail El-Naggar, Advisor

Prof. Waleed Khalil

Prof. Patrick Roblin
Abstract

The continual physical shrinking of semiconductor device dimensions is allowing for more integration between the previously segmented digital logic, memory, analog, and radio frequency domains – heralding the “More than Moore” era. Although able to meet the performance requirements for high-speed analog and RF, the devices are not guaranteed to always run at their typical sweet spot. The drifts from the optimal operation are due to many factors related to the silicon process and its response to changes in voltage and temperature, or what is collectively named $PVT$ (Process, Voltage, Temperature) variations. These variations are a problem in all the integrated domains of the chip; however, RF circuits fail, in a more disproportionate manner, at sustaining proper operation over PVT. This makes them more prone to performance degradations and loss of yield when fabricated, in contrast to digital chips that can achieve near perfect yield. Putting both RF and digital together on a single chip, the hybrid system obviously inherits the lower yield, negating all the integration advantages. Therefore, the RF portions, in a sense, represent the SoC’s Achilles’ heel; in essence, an overly powerful and densely integrated chip can be made useless by a smaller underperforming portion of the chip.

The ultimate goal is to increase the yield of the RF blocks by actively maintaining them in their optimal operating region. This proves to be a non-trivial task, as the
operating conditions of the system at all times need to be known. For complex integrated systems, full verification during fabrication testing is quite prohibitive, in time and cost. A solution would be to build self-testing, and eventually self-healing, systems. Built-in-Self-Test (BiST) paradigms have already established themselves in the validation of digital blocks but are now becoming an increasingly active domain of research and development in RF. The notion of migrating RF test functionality to inside the chip brings us one step closer to cognitive-like radios. If RF blocks and systems can test for, and extract, their performance, then the ability to calibrate and cancel discrepancies can also be built into the system. Hence, Built-in-Self-Calibration (BiSC) can be layered on top of BiST to result in auto-correcting RF impairments at the block and system levels.

In this dissertation, we discuss the problems set forth by increased integration and decreased circuit robustness. We also express the requirements for building efficient true self-test mechanisms using on-chip resources not only as value-added elements but also as necessary components for successful first-pass success of RF SoCs. An efficient RF sensor is presented along with the different possible built-in-tests for which it can be employed. The implementation of these on-chip test strategies aid in the development of calibration techniques that leverage the strengths of the more robust parts of the system to cover up the weaknesses of the others. As such, the digital domain is fully exploited to augment the capabilities of RF circuits by providing them with an added degree of tunability with the goal of enabling performance steering capabilities. The dissertation strives to present an approach and a description of a process that fits perfectly into the
premise of, and promise of, highly performing first-time-right design of RF SoC moving into the nanometer regimes.
To my loving family
Acknowledgments

This work was only possible with the direct support of wonderful mentors, many colleagues, friends, and family. I am truly blessed to have met them all through my journey.

First, I would like to extend my sincerest gratitude to Prof. Mohammad Ismail for giving me the opportunity to conduct my research at the Analog VLSI Lab. It has been a great 4 years under his supervision and guidance. My best wishes go to all the Lab’s members; past, present and future.

I would also like to deeply thank Prof. Waleed Khalil for his invaluable mentorship at the ElectroScience Lab. Prof. Khalil’s group has been a great group to collaborate with during the last phases of my studies.

In Columbus, my regards go to all the wonderful friends and loved ones.

From across the pond, I would like to thank the friends and colleagues in Sweden and Lebanon. I wish them all the success in their lives and endeavors.

Last but not least, I would like to dedicate this work to my wonderful family, my father Samir, mother Amal, brother Maroun, and sister Maria. It is to their love and belief in me that I owe my ability to persist forward.
Vita

October 6\textsuperscript{th}, 1983..............................................Born – Zalka, Lebanon

2005..........................................................B.E. Computer and Communication Engineering, American University of Beirut, Lebanon

2007..........................................................M.Sc. Electrical Engineering, Royal Institute of Technology (Kungliga Tekniska Hogskolan), Stockholm, Sweden

2007 to present ....................................Graduate Research Associate, Department of Electrical and Computer Engineering, The Ohio State University

Publications

Patents:

- Sleiman Bou-Sleiman and Mohammed Ismail El-Naggar, “Dynamic Self-Regulated Charge Pump Circuit with Improved Immunity to PVT Variations,” invention disclosure (Tech ID #11088) filed with The Ohio State University in November 2010
Books and Book Chapters:


Journal papers:


• Sleiman Bou-Sleiman, Mohammed Ismail, “Dynamic self-regulated charge pump circuit with improved immunity to PVT variations,” *submitted to the IEEE Transactions on Very Large Scale Integration (TVLSI)*

**Peer-reviewed conference papers:**


Fields of Study

Major Field: Electrical and Computer Engineering
# Table of Contents

Abstract .............................................................................................................................................. ii

Vita .................................................................................................................................................... vii

List of Tables ....................................................................................................................................... xv

List of Figures ...................................................................................................................................... xvi

Chapter 1 | Introduction and Motivation ............................................................................................ 1

1.1 The Need for Robust RF and mm-Wave ICs ................................................................................. 3

1.1.1 Integration Trends in CMOS .................................................................................................. 4

1.1.2 CMOS Scaling Effects .......................................................................................................... 7

1.1.3 Cost Factors ............................................................................................................................ 14

1.2 Aim and Scope of this Manuscript .............................................................................................. 17

Chapter 2 | Radio Systems Overview: Architecture, Performance, and Built-in-Test ................. 20

2.1 Transceiver Architectures .............................................................................................................. 21

2.1.1 Basic Communication System Architecture ........................................................................ 21

2.1.2 Heterodyne and Homodyne Configurations ......................................................................... 24

2.1.3 Quadrature Signal Processing ............................................................................................... 29
2.1.4 Transceiver Architecture for Multi-band Multi-standard SoCs .......... 31

2.2 RF System and Block Performance ................................................................. 32

2.2.1 System Metrics .................................................................................. 33

2.2.2 Component Metrics .......................................................................... 40

2.3 Integrated Radio and System-on-Chip Testing ............................................ 49

2.3.1 Built-in-Test Techniques .................................................................. 51

2.4 Summary ...................................................................................................... 56

Chapter 3 | Efficient Testing for RF SoCs ................................................................. 57

3.1 On-Chip Test Migration and Portability ....................................................... 58

3.2 A BiST-ready RF SoC .............................................................................. 62

3.3 RF Amplitude Detectors for RF BiST ......................................................... 65

3.3.1 Detector Requirements ...................................................................... 66

3.3.2 Detector Architectures ....................................................................... 68

3.3.3 Proposed Detector Design ................................................................ 69

3.3.4 Implementations for RF and mm-wave BiST ...................................... 75

3.4 Summary ...................................................................................................... 88

Chapter 4 | RF Built-in-Self-Test .......................................................................... 90

4.1 Specification-based Tests using the RF Amplitude Detector ....................... 90

4.1.1 Gain ..................................................................................................... 93
6.1.1 Analysis .................................................................................................................. 141

6.1.2 Proposed modulator ............................................................................................ 145

6.2 Dynamic Self-Regulated Charge Pump Circuit .................................................... 148

6.2.1 Proposed method and circuit ............................................................................. 149

6.2.2 Simulation results ............................................................................................... 153

6.3 Conclusions ........................................................................................................... 157

Chapter 7 | Conclusions ............................................................................................... 159

References ....................................................................................................................... 163
List of Tables

Table 1.1   CMOS Technology Roadmap ................................................................. 8
Table 1.2   Threshold voltage variability................................................................. 13
Table 3.1   Important transceiver parameters to measure ......................................... 64
Table 3.2   RF detectors in literature ................................................................. 68
Table 3.3   Comparison of the implemented RF amplitude detectors ....................... 88
Table 4.1   Test setups for the various RF blocks.................................................. 92
Table 4.2   Actual versus predicted parameters for the 60GHz LNA ...................... 109
Table 6.1   Optimal feedforward single-loop modulators and their configurations 144
Table 6.2   Resource usage comparison................................................................. 147
List of Figures

Fig. 1.1 System-on-Chip: single-chip radio systems with RF, analog, memory and digital ........................................................................................................................................ 1

Fig. 1.2 Data rates and ranges of various wireless standards.................................................. 6

Fig. 1.3 Transistor cut-off frequencies for different processes and geometry nodes........... 9

Fig. 1.4 Intrinsic and parasitic channel capacitance and resistance per technology node ........................................................................................................................................ 10

Fig. 1.5 The many contributors to variability in nanometer CMOS technologies.......... 12

Fig. 1.6 The effects of variability and tighter specifications and requirements on yield 13

Fig. 1.7 Moore's observation on the relative manufacturing cost and number of integrated components [1]........................................................................................................ 15

Fig. 1.8 The opposing trends in transistor cost and lithography tool cost [12]............. 15

Fig. 2.1 Basic communication system and its constituent blocks............................................ 22

Fig. 2.2 The mixer as a frequency conversion block ............................................................ 23

Fig. 2.3 Homodyne and heterodyne frequency conversion ............................................... 24

Fig. 2.4 The image problem and the need for image rejection ........................................... 26

Fig. 2.5 An example signal band with dc-free encoding ..................................................... 27

Fig. 2.6 Image problem in direct-conversion architectures ................................................. 28

Fig. 2.7 Constellation diagrams for various complex modulation schemes ................... 30

xvi
<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.8</td>
<td>Direct conversion transceiver architecture</td>
<td>32</td>
</tr>
<tr>
<td>2.9</td>
<td>BER versus SNR for various modulation schemes</td>
<td>34</td>
</tr>
<tr>
<td>2.10</td>
<td>IQ plane with ideal and measured symbol locations</td>
<td>35</td>
</tr>
<tr>
<td>2.11</td>
<td>Two-tone intermodulation spectrum</td>
<td>38</td>
</tr>
<tr>
<td>2.12</td>
<td>Linearity characteristics</td>
<td>39</td>
</tr>
<tr>
<td>2.13</td>
<td>Ideal and nonlinear amplifier effects</td>
<td>41</td>
</tr>
<tr>
<td>2.14</td>
<td>Nonideal mixer effects</td>
<td>43</td>
</tr>
<tr>
<td>2.15</td>
<td>PLL block diagram</td>
<td>44</td>
</tr>
<tr>
<td>2.16</td>
<td>Jitter and phase noise</td>
<td>46</td>
</tr>
<tr>
<td>2.17</td>
<td>Effects of phase noise on mixing</td>
<td>47</td>
</tr>
<tr>
<td>2.18</td>
<td>IQ demodulator: effects of amplitude and phase mismatch</td>
<td>48</td>
</tr>
<tr>
<td>2.19</td>
<td>Phase and amplitude imbalance effects on image rejection</td>
<td>49</td>
</tr>
<tr>
<td>2.20</td>
<td>Loopback testing configuration</td>
<td>53</td>
</tr>
<tr>
<td>2.21</td>
<td>Alternate testing: parameter, signature, and specification spaces and their mapping [9]</td>
<td>54</td>
</tr>
<tr>
<td>2.22</td>
<td>Digitally-assisted analog/RF circuit</td>
<td>55</td>
</tr>
<tr>
<td>3.1</td>
<td>A BiST-ready RF SoC</td>
<td>63</td>
</tr>
<tr>
<td>3.2</td>
<td>RF-to-dc curves for various detectors in literature</td>
<td>69</td>
</tr>
<tr>
<td>3.3</td>
<td>Proposed detector architecture</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>Proposed detector characteristics</td>
<td>72</td>
</tr>
<tr>
<td>3.5</td>
<td>One- and Two-tone transient responses with all tones having same amplitude</td>
<td>74</td>
</tr>
<tr>
<td>3.6</td>
<td>Pseudo-differential RF amplitude detector core</td>
<td>76</td>
</tr>
</tbody>
</table>
Fig. 3.7  Two-bit programmable voltage bias source .............................................. 77
Fig. 3.8  Response of microwave amplitude detector at 2.5GHz ................................. 78
Fig. 3.9  Combined continuous linear response of the detector for input frequencies
between 0.5 and 9 GHz.......................................................................................... 78
Fig. 3.10  Square-law and linear regions of the detector response at low RF signal
amplitudes.............................................................................................................. 79
Fig. 3.11  Deviations of the detector's response due to temperature and process
variations.................................................................................................................. 81
Fig. 3.12  Zero-RF re-referencing to increase measurement accuracy......................... 81
Fig. 3.13  mm-wave amplitude detector core and bias circuits ................................. 82
Fig. 3.14  mm-wave sub-ranged amplitude detector characteristics........................ 83
Fig. 3.15  Self-adjusting RF Amplitude detector....................................................... 85
Fig. 3.16  Self-adjusting RF detector modes under different temperature and process
variations.................................................................................................................. 87
Fig. 3.17  Maximum detection errors (Monte Carlo simulations) across the amplitude
range....................................................................................................................... 87
Fig. 4.1  BiST-ready SoC architecture with the important measurement points along the
transceiver.................................................................................................................. 91
Fig. 4.2  Setup for one-tone test for gain measurement ............................................ 94
Fig. 4.3  Two-tone test setup and IM3 measurement.................................................. 96
Fig. 4.4  Quadrature amplitude and phase mismatch measurement ......................... 98
Fig. 4.5  Downconversion mixer test setup for LO to RF port isolation ................. 100
Fig. 4.6 Upconversion mixer test setup for LO feedthrough .................................................. 101
Fig. 4.7 RF detector output and mapped output amplitude in response to a two-tone
input with varying IM3 component .......................................................................................... 104
Fig. 4.8 Predicted versus actual IM3 amplitude ........................................................................ 104
Fig. 4.9 Phase imbalance prediction versus the actual imbalance (left axis); prediction
error in degrees (right axis) ....................................................................................................... 106
Fig. 4.10 2.4GHz LNA gain extraction: actual versus predicted gain curve ......................... 107
Fig. 4.11 60GHz LNA used for mm-wave BiST ....................................................................... 108
Fig. 4.12 Effect of the mm-wave detectors on the LNA characteristics: with and without
.................................................................................................................................................. 108
Fig. 4.13 Actual and predicted gain curve after one-tone test sweep ................................. 110
Fig. 4.14 Third order intermodulation extraction and IIP3 measurement ............................. 110
Fig. 5.1 Self-calibration loop and its components ................................................................. 117
Fig. 5.2 Digitally-assisted RF and mm-wave circuits implemented with wide operation
flexibility and ranges always containing the optimal operation point ................................. 119
Fig. 5.3 Iterative calibration algorithm with changing PVT conditions .............................. 121
Fig. 5.4 LNA with digital calibration for input match and output load tuning ................... 124
Fig. 5.5 Transient snapshot of LNA calibration routine and the corresponding RF
amplitude detector dc output ...................................................................................................... 126
Fig. 5.6 Double-balanced CMOS mixer with digital tuning knobs for gain and linearity
calibration [8] .......................................................................................................................... 128
Fig. 5.7 Basic Farrow filter structure ..................................................................................... 131
Fig. 5.8 Mixed-mode IQ imbalance compensation ................................................................. 132
Fig. 5.9 Test tone demodulation: quadrature phase mismatch before and after compensation ......................................................................................................................... 132
Fig. 6.1 Output phase noise due to Sigma-Delta modulation: contribution from the modulator and the charge pump static gain mismatch (example spectra shown for a third order SDM with a noise shaping of 60db/decade) .............................................. 138
Fig. 6.2 Feed-forward single-loop SD Modulator ................................................................. 139
Fig. 6.3 Plot of stable B1 and B2 coefficient ranges for B3=0.5 and integrator configuration of DNN. Digitally implementable coefficient pairs are denoted by circles ................................................................................................................................. 141
Fig. 6.4 Integrated phase noise performance of three 3rd order modulators as a function of charge pump mismatch (Fref=40MHz, loop bandwidth=1MHz) ............... 143
Fig. 6.5 Optimal 3rd (left) and 4th (right) order modulators as a function of Fref, loop bandwidth, and charge pump mismatch percentage .................................................. 144
Fig. 6.6 Proposed multimode reconfigurable ΣΔM architecture ....................................... 147
Fig. 6.7 Output spectra of third and fourth order modulators (a)-(f). The theoretical noise transfer function for each modulator is plotted (in dotted line) on top of the PSD [vertical units: db/bin, horizontal: Hz] ................................................................. 148
Fig. 6.8 Conventional charge pump and current characteristics ........................................ 150
Fig. 6.9 Proposed charge pump circuit with dual-feedback .................................................. 151
Fig. 6.10 Feedback concept of the proposed charge pump .................................................... 152
Fig. 6.11 Proposed CP current matching across output voltage versus a typical CP .. 154
Fig. 6.12 Current mismatch with and without replica compensation .......................... 155
Fig. 6.13 Monte Carlo simulation results for current mismatch at T=30 ................... 156
Fig. 6.14 CP mismatch across PVT ........................................................................... 157
Chapter 1 | Introduction and Motivation

After having established itself as the dominant process for digital microelectronics, the past decade saw the emergence of CMOS as a viable process for Radio Frequency (RF) blocks. The advances in semiconductor technologies allowed for new long-sought-after opportunities for high scale integration of complete systems on a single chip, or System-on-Chip (SoC). The ability to successfully embed RF, analog, and digital portions of radio systems into a fully integrated solution brought about many new and interesting products and applications (Fig. 1.1). As much as this has opened up possibilities, such high integration is not without its shortcomings.

Fig. 1.1 System-on-Chip: single-chip radio systems with RF, analog, memory and digital
With every new boundary overcome in producing the next generation of advanced process technologies, we begin to observe a new set of challenges and impediments. This is more visible with analog and RF circuits. Inasmuch as digital circuitry and memories benefit from miniaturization, analog and RF circuits’ behavior starts to show increased susceptibility to deviate from optimal performance, therefore negating any gains made possible by their smaller sizes. This causes analog and RF circuits to suffer from low yield due to process, supply, and temperature (PVT) variations and therefore require several expensive silicon design and manufacturing cycles to meet their specifications. Also, when co-existing with a mass of fast-switching digital logic on the same substrate, they are inherently subjected to additional noise coupling. These issues have put increased limitations and restrictions on the design of RFIC blocks in platform baseband SoCs. To tackle the problems associated with process shifts, parasitic elements, and changing operating conditions, recent efforts concentrating on novel design techniques have emerged with the goal of minimizing the yield loss in radio SoCs. To the industry, this essentially translates to reduced engineering costs, faster product development, and faster time to market.

To reduce yield loss due to variability, post-silicon calibration is necessary to compensate the performance degradation. Calibration of RF blocks, however, is not an easy and straightforward task – it requires enhanced test mechanisms to be incorporated into the design of the system. Responding to these difficulties, Design for Testability (DfT) within RF systems has caught ground and enabled the implementation of Built-in-Self-Test, or BiST. This emerging field of research and development opened the way for
circuit self-awareness. Self-aware circuits will ideally not only be able to test themselves – detecting their operating conditions – but also correct for deviations through self-calibration procedures, to revert to optimal performance under any operating point. This vision of self-healing RF circuits can then only be achieved with new methodologies for pre-silicon designs incorporating robustness enhancement at their center.

In this chapter, we briefly describe the factors enabling CMOS RFICs and radio SoCs. We discuss the effects of technology scaling, variability in the nanometer regime, and some of the costs associated with integrated circuits. This will provide scope to the rest of the manuscript which aims at presenting a description of the challenges in design and test (Chapter 2) as well as suggesting solutions towards building robust first-pass RF SoCs through efficient self-test (Chapters 3 and 4) and self-calibration (Chapter 5).

1.1 The Need for Robust RF and mm-Wave ICs

As nanoscale CMOS pushed its operating boundaries beyond the digital into the high-frequency analog and RF and more recently the mm-wave realm, the focus has been on how to increase the yield of these ICs, standalone and in heterogeneous systems. The ultimate goal is then to achieve first-pass success through first-time-right design techniques and methods for radio SoCs. To understand the need for building robust ICs, we briefly discuss the driving forces and enabling factors for CMOS radios in terms of integration trends and device scaling. Like most technologies, highly-integrated CMOS SoCs will only make sense if they provide an economically viable solution. For that, the
cost implications of the entire cycle spanning design, manufacturing, and test plays into the overall equation.

1.1.1 Integration Trends in CMOS

In 1965, Intel cofounder Dr. Gordon E. Moore penned a formulation describing his observations regarding the near-exponential increase in semiconductor component integration and extrapolation of the trend into the then forthcoming decade. His statement, Moore’s law, as it came to be called, is in essence a remark on the cost of integration: “The complexity for minimum component costs has increased at a rate of roughly a factor of two per year” [1]. Revised later to the more common notion of doubling transistor densities every two years, the trend held up for over half a century with effects not only in the density of transistors but also in related computational technologies, such as memory capacities, processing power, image sensors, and others.

Semiconductor companies have adopted Moore’s law as a benchmark and milestone in their technology advancements and goals. This has spawned novel technologies and sparked innovations in trying to keep with the ever increasing exponential goals. When photolithography techniques could not surpass the limitations of light wavelengths to create smaller and smaller transistors, breakthroughs in extreme ultraviolet (EUV), double-patterning, and maskless techniques have ensured the survival of the trend to minimum sizes down to 22nm and 14nm at the time of writing [2][3]. Back in 1971, the first general purpose microprocessor, the Intel 4004, held approximately 2,300 transistors at 10µm length with processing power around 0.07 Million Instructions per Second
(MIPS). In 2011, Intel’s own Core i7 processor counts 915 million transistors at 32nm length capable of close to 160,000 MIPS. Today’s systems continue the extreme scaling trend and are expected to increase their densities and computational capacities in the near future, albeit with more innovations as we close in on physical atomic limitations.

As people’s mobility increased so did their demand that computational powers travel with them. The surge in integrated electronics has catalyzed yet another revolution: the wireless revolution. From pagers in the mid 1970s to the 80s and 90s’ mobile phones then notebook computers and more recently smart-phones and also stretching to myriad other applications such as body-implantable wireless medical devices or even smart utility metering, the wireless revolution enabled prompt and seamless communication and made ubiquitous connectivity a staple of our daily lives. Fig. 1.2 shows a subset of the prevalent wireless standards, their data rates and ranges, showcasing the breadth of applications from near field connectivity to long distance 3G and 4G wireless communications.

Digital computation and wireless communication have then merged to create single integrated systems for data transmission. The successful integration of the digital and radio frequency (RF) capabilities in plain silicon CMOS, a process heavily focusing on digital, made building platform System-on-Chips a very attractive and viable option. These heterogeneous systems include circuits from different signal processing domains – RF, analog and digital. As such, the design, test, and manufacturing expertise previously confined to each of these types of circuits, has to merge accordingly presenting new challenges and possibilities.
CMOS’ circuit boundaries have been able to migrate from the digital to the high-speed analog and RF due to the effects of physical scaling. During the past decade or so, transceiver blocks started appearing in CMOS as the process was able of operation in the RF spectrum (0.4 – 10 GHz). One of the last blocks to enter the CMOS realm is the Power Amplifier (PA) and with it added to the mix, little remains in terms of transceiver blocks that cannot be integrated on the same silicon substrate. Single-chip solutions for popular standards such as WiFi, WiMAX, and LTE are now available and are designed to meet the growing demands for high speed data, enhanced applications, and high levels of mobility.

More recently, as higher speeds are achievable in CMOS, the mm-wave spectrum has gained interest for commercial exploration, especially for short-range high-
throughput applications. While CMOS has never been the process of choice for extreme high frequencies, its promise for increased integration and reduced cost make it the strongest contender to cover non-niche applications steered towards the mass consumer markets, e.g. WirelessHD (IEEE 802.15.3c) [4].

1.1.2 CMOS Scaling Effects

The effects of scaling were first observed by the early researchers in the semiconductor field where they noticed benefits in nearly all aspects of a shrunken transistor. Physical process scaling has steadily continued to bring in advantages in speed and chip area – two very sought after metrics for future investment in any technology. However, disadvantages also abound, more so with the newer scaling efforts as further size reductions are now less possible.

Table 1.1 shows the more recent and expected CMOS scaling trends [5]. As the transistor’s length is decreased, we see an inversely proportional increase in the speed of the MOS device ($1/L_{gate}$). The reduction in area, on the other hand, is proportional to $L_{gate}^2$. Supply voltages are also reduced to lower the power consumption coupled with a decrease in the threshold voltages ($V_{th}$). This reduction in threshold voltage is necessary for proper operation under lower voltage supplies but has the disadvantage of exponentially increasing the subthreshold leakage, which is starting to account for 50% of the total power consumed in modern embedded systems [5]. As further scaling of planar CMOS is unsustainable, new transistor structures are under study for the 14nm nodes and beyond, such as FinFETs [7] and three-dimensional tri-gate structures [8].
With each advanced process node, faster transistors became possible that not only increased the digital circuits’ operating frequency but also made it possible to build RF and mm-wave circuits spanning the low to tens of GHz range. While the latter types of circuits were typically manufactured in SiGe and GaAs due to their superior RF performance (noise, speed, etc...), CMOS device $f_T$ has steadily increased with each generation making it capable of handling high-frequency wireless. Fig. 1.3 compares silicon CMOS with the other process technologies that occupy the RF and mm-wave space. CMOS has always led the pack in scaling with the other technologies trailing by approximately two generations [9]. For a process technology to be capable of a certain application, its $f_T$ should exceed the operating frequency by tenfold, for small-signal operation, or twice, for oscillation and power transistors. As seen in Fig. 1.3, nanoscale CMOS is well within the RF spectrum and closing in on the mm-wave range with each process generation.
For CMOS to operate RF devices, it needs to exhibit certain figures of merit such as gain, noise figure, output power, and linearity among others. As many of the changes to bulk CMOS are done to implement faster digital circuitry, this does not necessarily translate to better analog and RF metrics. Variations are much more pronounced in deep-submicron and nanometer devices thereby manifesting in increased levels of RF distortion and loss of performance [10][11]. Analog and RF circuits depend a great deal on matching between devices, e.g. differential pairs. Mismatches due to variations in gate dielectric, random dopant fluctuation, and line-edge/line-width roughness, weigh in on several aspects of the device’s behavior, such as sub-threshold currents and threshold voltages [13]. Undesired parasitics that were once too small to notice are now sufficiently
large with respect to the intrinsic device parameters – and are expected to increase further, as shown in Fig. 1.4 for channel resistance and capacitance [12].

![Intrinsic and parasitic channel capacitance and resistance per technology node](image)

**Fig. 1.4** Intrinsic and parasitic channel capacitance and resistance per technology node

Another type of parasitics, interconnect parasitics, plays a pivotal role in the performance of transceivers and SoCs, especially when routing high frequency signals, such as RF and high-speed I/O. With highly integrated RF SoCs, the number of interconnect layers has increased to allow for sufficient connectivity for the embedded complex circuitry – typical submicron processes can have up ten metal layers, with top metals suitable for RF passives and signal routing. So, apart from the device parasitics that limit the performance of individual circuits, SoC performance is also limited by the interconnect between, and beyond of, the individual blocks. These limitations arise from factors such as RC delay, IR drop, and cross-talk [14]. With parasitic effects amplified at very high frequencies such as in mm-wave applications, metal wires’ passives like sheet
resistance, coupled capacitance to nearby metals and substrate, and mutual inductance have to be embedded and modeled into the design process [15][16]. Their effects on system parameters cannot be ignored, especially with heterogeneous systems having fast switching digital circuits and precision analog and RF on the same silicon substrate [17].

Process variations and parasitics skew the performance of analog circuits and force designers to include substantial margins to reduce and try to eliminate yield loss. Over-design usually comes at the expense of non-optimal power solutions, paramount to over-kill. Fig. 1.5 shows the various forces affecting circuit design and reducing yield with their adverse variability. It is quite important to have accurate models that capture these effects at early stages of the design as well as precise control over passives. However, designers have to take special care and invest more design effort to ensure performance robustness from the RF and analog parts. Apart from the intrinsic changes in the process, environmental variations in the operating conditions, such as temperature and supply voltage changes, have an equally important effect on the circuit behavior. Collectively called PVT variations (for process, voltage, and temperature), they form a set of corner cases which designers have to account for. NMOS and PMOS devices can either be fast or slow, resistance and capacitance values can vary by several tens of percentage points, and each device has its own temperature gradient – any combination of these cases and others form a set of corners that can manifest in a silicon run, even within the same wafer. With newer processes, the number of possible corner cases has dramatically increased forcing designers to resort to substantial over-design. Therefore, solutions are needed to mitigate performance degradations through novel techniques in process,
design, and layout. And in order to reduce the silicon spins and achieve first-time-right design, these best-practice techniques can be augmented with self-test and self-calibration schemes – both topics discussed in later chapters.

![Diagram of factors contributing to variability in nanometer CMOS technologies](image)

**Fig. 1.5** The many contributors to variability in nanometer CMOS technologies

The advantages enabled by scaling can be easily shadowed by the disadvantages coming from the increased variability. Table 1.2 lists the increasing threshold voltage variations with decreasing geometries [18]. The superposition of several such variations leads to a very wide distribution of device performances, to the first degree, and eventually circuit/block performance uncertainty. Moreover, as we look at newer wireless standards, they impose more stringent system performance requirements effectively narrowing the pass/fail boundaries. These complimentary effects result in more yield losses (Fig. 1.6) and subsequently more design, test, and fabrication investments.
Digital designs have so far been able to keep a very high yield, owing to their discrete nature of operation, something RF and analog blocks cannot boast. The loss of performance from the latter blocks will eventually set the pass/fail limit of the entire chip – rendering completely flawless digital designs useless with a single faulty, or sub-performing, RF block in the SoC.

<table>
<thead>
<tr>
<th>Technology</th>
<th>180nm</th>
<th>130nm</th>
<th>90nm</th>
<th>65nm</th>
<th>45nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma V_{th}$</td>
<td>5.8%</td>
<td>8.2%</td>
<td>9.3%</td>
<td>10.7%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Table 1.2 Threshold voltage variability

![Performance distribution diagram](image)

Fig. 1.6 The effects of variability and tighter specifications and requirements on yield
1.1.3 Cost Factors

As mentioned previously, Moore’s law’s original statement is an observation regarding cost. Fig. 1.7, Moore’s original redrawn sketch from [1], shows that the cost of integrating a single transistor was decreasing with time. And ever since, the number of transistors that can be integrated at minimum cost has been increasing exponentially making the cost of a single transistor extremely cheap, as seen in Fig. 1.8. However as the latter figure also shows, silicon foundry tool costs are on the increase translating to more expensive design fabrication for the IC firms. Driving the costs up is the foundries’ continued investment in research and development to push the limits of the technology, helping them create larger wafers, increasing the yield, and integrating more devices onto silicon [19]. All these goals attempt to decrease the cost of integrating a transistor – however, the probability of that transistor being a faulty device is increasing with each denser process node. This necessitates proper and accurate device testing at the production level. Testing costs have generally remained constant but with the decrease in integration costs, testing cost per transistor has become comparable to the latter. To the IC design firms, fabrication costs are quite expensive with typical nanometer CMOS mask sets cost reaching a couple million US dollars [20].
Fabrication cost is but one of the many costs associated with the lifetime of an IC product. Prior to mass production, several iterations of design, manufacturing, and testing might be necessary to reach the desired performance of an SoC. All of these phases entail sizeable costs. A single design cycle can take upwards of one year, between setting system specifications, system design, then circuit design, layout, and post-layout
verification – with multiple possible revisions to any of the previous. During that primary design cycle, costs include salaries, software and design tools. For any subsequent refinement cycle, non-recurrent engineering (NRE) costs start chipping from the possible revenue, in actual cost and time. Of extreme importance in a very competitive semiconductor market is the value of time. A balance needs to exist between reducing any delays in time-to-market to maximize profits over the product’s life cycle, and rushing that product to the market without proper verification and testing. Defective products that reach the end customer can cost more than 10,000 times their selling price to replace [18].

Testing RF SoCs is not a trivial task given the complex nature of these systems. Based on the issues discussed so far, it becomes evident that RF SoC testing needs to achieve high accuracy while reducing costs. Testing costs are in themselves also high, both at the prototype and mass production stages. In the first, IC design firms use laboratory setups with a handful of expensive test equipment performing multiple types of tests to assess the performance of the designed chips. At production, a more automated testing solution in the form of Automatic Test Equipment (ATE) exists to quickly verify proper system functionality. In either cases, RF SoCs present a challenge in testing as they include a large number of test points and possible faults that are not easily accessible from the outside. To this end, circuit designers have recently resorted to embed test-enablers and even complete test macros on-chip in addition to the main system blocks. This is the basis for the Design-for-Testability (DfT) approach. With DfT, some of the testing can be migrated on-chip and test results can be interfaced and read from off-chip
using low cost equipment. Embedded testing is referred to as Built-in-Self-Test (BiST) whereby a system tests itself using on-chip resources rather than relying on the outside for all the testing. Although quite mature for digital testing, BiST for RF is still under active research and development. In RF SoCs, on-chip resources are plenty and hence capable of a number of test-specific functions, such as test signal generation and test result analysis. Hence, RF BiST can leverage the processing power of the digital core of the SoC to perform system- and even block-level tests with the goal of reducing test costs. Chapter 2 will present more on the various built-in testing paradigms for RF SoCs.

1.2 Aim and Scope of this Manuscript

Nanometer CMOS has opened the door for a new horizon and possibilities for highly integrated and powerful wireless computing. The vision of ubiquitous access to data and the ability to utilize this data for countless applications is one of great interest to consumers, companies, and academia.

But for these technologies to reach maturity they have to surpass many challenges in their design, manufacturing and test. Only when cost-effective can a certain application make its way to the mass market. Therefore, IC designers have to overcome many of the limitations that prevent the successive push towards more integration of computing powers and wireless connectivity. As the current state of process technologies allows for such capabilities to be implemented, designers are going to lengths to make coexistence of powerful data processing and wireless data communication stable, robust, and effective, in in-field operation and more importantly in pre-deployment cost. Chapter 2
describes the building blocks of radio transceivers, system and block metrics, and the more suitable architectures for SoCs. More importantly, we also discuss the shortcomings and limitations of the radio system’s building blocks and describe built-in test techniques to quantify their performance.

The challenge then is how to benefit from the advances in various design and test domains to achieve first-time-right RF and mm-wave radio SoCs. As we move to atomic scales, robustness is severely degraded and has to be artificially augmented. This augmentation is only possible with novel designs built around test and calibration. Novel methodologies encompassing state-of-the-art techniques and circuitry are needed. The ultimate goal is to instill a heightened sense of cognition, along the lines of artificial intelligence, into radios and systems – steering their subpar performance up to specifications through the clever use of available resources. Flexible radios and smart systems will have to embed built-in-self-test and built-in-self-calibration (BiST and BiSC).

Design for testability and tunability is a hot field in research and development. The enabling circuits for testability are still being developed for RF and mm-wave; to this end, we describe one such effective and low-overhead circuit in Chapter 3. Many challenges exist in making on-chip detection reliably accurate, especially when dealing with very high frequency signals. Self-test examples are investigated in Chapter 4. Coupled with the clever use of robust resources to heal the less immune circuits through tunability, Chapter 5 discusses the use of analog and digital calibration to enable first-
pass silicon success, cutting costs and eventually freeing funds for further technological development.
Chapter 2 | Radio Systems Overview: Architecture, Performance, and Built-in-Test

In this chapter, we take a look at RF systems from the system to the block level. The overview presents an introduction to the basic communication system, its various architectures and implementations. An architecture best suited for SoC integration is highlighted. Throughout the Manuscript, elements will be added to that baseline configuration to enable and demonstrate true built-in self-test and self-calibration capabilities.

Section 2.1 starts with a description of a high-level basic communication system, the heterodyne and homodyne architectures, and their respective considerations. The more suitable architecture for SoC integration is then described. Furthermore, in Section 2.2, the important blocks in the radio system are highlighted along with their metrics and critical impairments. The different Built-in-Testing schemes for RF systems are presented in Section 2.3.
2.1 Transceiver Architectures

Wireless receivers and transmitters - called transceivers when they co-exist in a system - are often symmetrical and parallel in construction but accomplish opposing translations. They both contain elements that perform signal amplification, conditioning, frequency and type (analog versus digital) conversion. Ultimately, they perform the translation from and to wireless high-frequency analog signals and digitally-encoded data bitstreams.

In this section, a basic communication system architecture is first described and then its various configurations briefly highlighted. The different aspects of these configurations are explained and contrasted with their advantages and disadvantages – while highlighting the more suitable setup for integration in RF SoCs. Following that, the important radio system and block metrics are presented along with their impairments. Lastly, we describe the testing of transceivers as it pertains to highly integrated SoCs with primary emphasis on built-in-testability.

2.1.1 Basic Communication System Architecture

Fig. 2.1 shows a high-level diagram of a radio system: as shown, receivers operate unidirectionally from left to right, receiving a radio-frequency weak signal from the antenna and propagating that signal along the chain while performing conditioning (filtering), amplification, followed by downconversion of the radio-frequency content to frequencies suitable for digitization. The transmitter is at its simplest the mirror opposite.
The signal conditioning block between the antenna and the amplification block is usually a band pass filter (BPF) that allows through the required frequency while attenuating unwanted out-of-band frequencies. It is very common that this block’s functionality is merged into either the antenna (by the inherent nature of its design) or into the amplification stage (e.g. narrowband amplifiers). The amplification blocks in the receiver and the transmitter differ in their design optimizations. In the former, it is important that the amplifier introduces as little noise to the weak signal while amplifying it. In the transmitter, the amplifier is optimized to send a powerful clean signal out through the antenna. As such, the amplifier in the receiver is called a Low Noise Amplifier (LNA) whereas that in the transmitter, a Power Amplifier (PA). Depending on the specifications, multiple stages of amplification might be necessary.

The frequency conversion block comprises of a mixer coupled with a local oscillator (LO), or a cascade of such. Time-domain multiplication of signals, i.e. mixing, results in frequency-domain conversions. The frequency conversion factor is then determined by the oscillator frequency. It is imperative to note that signal mixing produces sideband
copies, located at the sum and difference of the respective signals (Fig. 2.2). Proper filtering and amplification are needed to pick out one sideband to propagate the chain. In a receiver, high frequency inputs are downconverted to low frequencies for easier handling by the following processing blocks. Downconversion is then the result of mixing the signal with the LO frequency followed by appropriate filtering to obtain the lower sideband. On the transmit side, low frequency signals supplied by the processor are upconverted to radio frequencies for transmission.

On either sides of the radio chain, a signal medium and type translation is necessary. On one end, the antenna acts as the air interface to convert between wired and wireless signals and on the other end, a signal type conversion block interfaces the analog and digital domains: Analog-to-Digital converters (ADC) on the receiver side and Digital-to-Analog converters (DAC) on the transmit side.
2.1.2 *Heterodyne and Homodyne Configurations*

Transceivers can be grouped into two major architectures, heterodyne and homodyne, differing mainly in the number of frequency conversions performed. In essence, the previously discussed high-level system architecture in Fig. 2.1 remains the same with the exception of whether the frequency translation is done in a single shot or through multiple conversions, as depicted in Fig. 2.3.

![Diagram of Homodyne and Heterodyne Frequency Conversion](image)

**Fig. 2.3 Homodyne and heterodyne frequency conversion**

Heterodyne systems use the latter, and most commonly in two steps. These frequency conversions shift the main signal to an intermediate frequency (IF) first, and then to the desired final frequency, whether it being RF towards the antenna or low-
frequency towards the processor. This segmentation allows for more system adaptability and gives the heterodyne systems superior selectivity and sensitivity. However, these advantages come at the expense of bulkier, more complex, and higher power-consumption circuits. Apart from the area and design overhead for two mixers and LOs, multiple mixing operations demand stringent filtering requirements. For one, band-pass filters operating around IF are needed in-between the two stages for channel selection. However, of more importance is filtering out the “image”.

The LO is usually a sinusoidal tone whose frequency contents are two impulses at symmetrically opposite frequencies around dc – positive and negative (as is the case for real signals). This causes the negative frequency tone to also contribute to the mixing. Therefore, signals located at frequencies symmetrical to the desired signal’s frequency with respect to the LO tone also exit the mixer at the same IF. To prevent this unwanted band (i.e. “image”) from overlapping and corrupting the intended signal, image reject (IR) filters are required. Fig. 2.4 depicts the image problem with a desired band at RF frequencies getting corrupted by an image signal when downconverted with and without an IR filter. High quality factor IR filters are often bulky and cannot be integrated in silicon, thus requiring additional routing of high-frequency signals off-chip and back. Therefore, for integrated solutions and System-on-Chip applications, a heterodyne receiver is seldom used.
In contrast, the single-shot conversion scheme is called the homodyne architecture. This architecture was originally developed to overcome the image problems in heterodyne systems through direct conversion to – and from – dc or baseband (BB), lending them the name zero-IF. This relieves the filtering requirements and results in less components thus making this type of architecture very suitable for SoC integration. However, homodyne architectures inherit their own set of drawbacks. For one, they require very stable and accurate frequency translation, with LO frequencies as high as the target frequency. Moreover, they also suffer from some problematic system and circuit issues that manifest themselves at dc and low frequencies. Nonidealities in the analog circuitry and even self-mixing of the LO signal due to poor mixer isolation result in dc-

Fig. 2.4 The image problem and the need for image rejection
offsets and low-frequency harmonics that eventually corrupt the desired signal itself. These effects are highlighted in the upcoming sections.

Moreover, low-frequency noise such as device flicker noise also significantly degrades the target signal. Recent advances in data coding and complex modulation schemes allow for dc-free signals that can be received without distortion due to self-mixing and other dc-offsets. For example, a typical WiMAX and LTE baseband (Fig. 2.5) does not use a dc sub-carrier therefore facilitating dc-offset removal without affecting the received signal.

Fig. 2.5 An example signal band with dc-free encoding

An intermediary architecture is the low-IF topology. It is a homodyne architecture in construction, but conversion is done to and from a low frequency IF. The dc-offset, LO feedthrough, noise and various nonlinearities are then of less concern as they do not affect the signal at low IF. On the other hand, direct digitization of the low-IF signal demands high sampling rate ADCs in the receiver and DACs in the transmitter. Also, the image folding still exists at low-IF but its rejection can be migrated to the digital domain.
Even though direct conversion promises image free modulation and demodulation, technically this is not completely true. In zero-IF systems, an image still exists: the signal itself. Apart from amplitude modulated (AM) signals, the upper and lower sidelobes of a certain band of interest are not necessarily symmetrical, as is the case for frequency and phase modulation. Therefore, upon conversion to dc, the desired band will overlap with its mirrored self, i.e. the distinct upper and lower sidelobes now overlap and corrupt each other (Fig. 2.6). The same phenomenon happens at the second and final conversion of a heterodyne system. To overcome this limitation, a phase and frequency decoupling mechanism is needed – hence enters quadrature mixing.

Fig. 2.6 Image problem in direct-conversion architectures
2.1.3 Quadrature Signal Processing

In zero-IF as well as in the last stage of a heterodyne, the mirror-copy image frequency overlaps with the target frequency band and cannot be solved with filtering. In low-IF scenarios, the image is very close to the desired band thus requiring a very sharp IR filter. In these cases, quadrature mixing can be used to separate the positive and negative frequencies. This is achieved by orthogonal mixing, with in-phase (I) and out-of-phase (Q) components, 90° apart – or in other words, a cosine and sine. This creates two parallel signal paths in which the incoming or outgoing signals are mixed with orthogonal sinusoids generating quadrature signals with phase shift. The phase shift results in interesting properties upon recombination where the image adds destructively while the target frequency adds constructively.

Also an increasing number of today’s digital communication standards rely on phase modulation and therefore have quadrature aspects as the basis of their schemes. In these standards, data is encoded and decoded as a complex signal with the in-phase and out-of-phase components as the real and imaginary parts, respectively. This forms an IQ space resembling a two-dimensional Cartesian space commonly referred to as a signal constellation plot or diagram. Fig. 2.7 shows the ideal locations of several such schemes where data can take on one of these symbols differentiated by phase shifts (Phase Shift Keying, PSK) or combined amplitude and phase shifts (Quadrature Amplitude Modulation, QAM). For high data rate applications, very dense constellation spaces are needed to encode and decode an increasing number of bits per symbol, for example 64-QAM.
As an image rejection technique, quadrature processing is a very attractive option. However, sufficient image rejection is only guaranteed when the two paths are matched. Any mismatch in the dual paths will result in incomplete cancelation of the image. In the case of digital information carried on the RF carrier, incorrect reception of phase and amplitude might result in interpreting an erroneous symbol, i.e. the real symbol constellation becomes distorted due to the $I$ and $Q$ images. The two mismatch mechanisms here are mainly due to gain variations (amplitude mismatch) or slightly out of quadrature LO (phase mismatch). Appropriate sideband rejection is hence a function of how matched the two paths are, for which a figure of merit is the Image Rejection Ratio (IRR). In the next section describing RF impediments, it will be shown how minor mismatches in either amplitude or phase rapidly deteriorate the IRR.

![Fig. 2.7 Constellation diagrams for various complex modulation schemes](image_url)
2.1.4 Transceiver Architecture for Multi-band Multi-standard SoCs

For integrated SoCs with multi-standard capabilities, a flexible transceiver architecture is highly desired. The most appropriate selection would then be the direct conversion architecture – zero- and low-IF: good performance from state-of-the-art LO (frequency synthesizers) and IQ modulators, fewer filters (better wide-band capabilities) and less power, less mixing product spurs, and for the case of newer standards, no active subcarriers at dc.

Depending on the communication system, a transmitter and receiver can function at the same time, given they operate at different frequencies, or alternatively use the same frequency but alternate in operation. These two scenarios are called frequency or time division duplexing, FDD and TDD, respectively. This allows both to share the same antenna while placing a routing block, usually a transmit/receive switch (time-domain splitting) or duplexer (frequency-domain splitting) usually as discrete off-chip components. A typical TDD direct conversion transceiver is shown in Fig. 2.8. TDD systems require only a single LO as the transmitter and receiver use the same frequency band but are duplexed in the time-domain through an RF-switch to and from the antenna. FDD systems on the other hand use different frequencies for the transmit and receive and therefore require two LO signals. In Fig. 2.8, the receiver is composed of a Low-Noise Amplifier (LNA) amplifying the signal with little added noise before being down-converted through two mixers driven by quadrature LO signals. The baseband signal can then be filtered using low pass filters before being amplified by Variable Gain Amplifiers (VGAs) to appropriate ADC levels. The transmitter chain, on the other hand, starts off
with the generation of I and Q analog baseband signals from the DACs to be low-pass filtered (to remove sampling frequency aliasing) and upconverted using an IQ modulator. The upconverted signal is further amplified for transmission by means of a Power Amplifier (PA), which might be preceded by an RF preamp.

![Direct conversion transceiver architecture](image)

**Fig. 2.8** Direct conversion transceiver architecture

### 2.2 RF System and Block Performance

For a complete system to work, the sum of its parts needs to offer good performance. On the macro level, system level metrics give insight to whether a radio meets specifications as demanded by a communication standard. On the micro level, the chainlike cascade of transceiver blocks puts requirements on each individual block to properly process its inputs and present a suitable output to the next block. Link budgets
are used to segment the performances to the various components in the system. These components have their own set of important metrics and have to meet these performance goals for the system to maintain the required overall performance. In the following, we highlight the most important metrics on the system and component levels and trace their dependence on circuit impairments.

### 2.2.1 System Metrics

Wireless standards often specify certain high-level metrics to ensure transmission quality. Looking at an end-to-end wireless system, it makes sense that the ultimate goal is the transmission and reception of error-free data. Therefore, system metrics such as the bit error rate and error vector magnitude are used extensively to test for standard compliance.

#### 2.2.1.1 BER and EVM

The Bit Error Rate (BER) is a definitive test of the performance of a communication channel. It is defined as the number of bits received in error divided by the total number of bits transmitted during a unit time. As such, BER can represent the end-to-end system performance, from modulation, transmission, propagation, reception and demodulation. Depending on the type of modulation, BER can behave differently with different values of the Signal-to-Noise Ratio (SNR) – the latter being the power of the signal in relation to the system total noise. In general, an increase in SNR causes a decrease in BER as signals
are clearly distinguishable from noise and correctly interpreted. Fig. 2.9 shows BER versus SNR for a number of modulation schemes showing the quick roll-off with respect to SNR.

![Graph showing BER versus SNR for various modulation schemes]

**Fig. 2.9** BER versus SNR for various modulation schemes

Another metric is the Error Vector Magnitude (EVM). As the digital data is transmitted or received, the actual symbol location, as mapped on an IQ constellation, might differ from the ideal location. EVM represents a measure of this discrepancy and hence provides an indication of modulator/demodulator performance. EVM provides a compacted reading of many parameters that affect the individual blocks of the system, such as poor IRR, phase noise, and carrier leakage (all of which are described in the coming section). The error is computed as the difference vector between the measured and ideal symbol. A graphical representation is shown in Fig. 2.10, where $\vec{v}$ is the vector...
denoting the ideal symbol location, \( w \) the measured symbol location, \( \theta \) the phase error, 
\(|w| - |v|\) is the magnitude error, and \( e \) is the error vector. The error vector magnitude is then defined as \(|e|/|v|\), and sometimes referred in percentage by dividing it by \(|v|\). The root-mean-square (RMS) EVM and phase error are then used to determine the EVM measurement over a window of several demodulated symbols.

![IQ plane with ideal and measured symbol locations](image)

Fig. 2.10 IQ plane with ideal and measured symbol locations

To remove the dependence on system gain distribution, EVM is normalized by \(|v|\), which is expressed as a percentage, or as a root-mean-squared over a measurement window.

BER and EVM testing give a high level reading of the system performance but both take a relatively long time in order to achieve a certain confidence level [1]. EVM testing
usually requires a shorter time and therefore studies have tried to relate BER to EVM for
various modulation schemes in an effort to reduce the test time [22].

2.2.1.2 Link Budget Analysis

A link budget study is a very important system level design step that analyzes the
required performance metrics for any given standard then distributes and divides the
system requirements to the constituent blocks in the RF receiver and transmitter chains.
System parameters of interest in a radio system are noise figure, gain, and nonlinearity.

An important performance metric is the Noise Factor \( F \) which measures the Signal-
to-Noise ratio (SNR) degradation from the input of the receiver to its output. The Noise
Factor is more often than not described in its logarithmic equivalent the Noise Figure
\( NF = 10 \log(F) \).

Each block within a transceiver is characterized with its individual noise factor \( F_i \) (or
noise figure, \( NF_i \)) and gain \( G_i \). Since a typical receiver or transmitter are composed of
cascaded blocks, the total gain is simply the product of the individual gains
\( G_T = G_1 G_2 ... G_n \) however the total noise factor can be described as follows:

\[
F_T = 1 + \left( F_1 - 1 \right) + \frac{(F_2 - 1)}{G_1} + \frac{(F_3 - 1)}{G_1 G_2} + ... + \frac{(F_n - 1)}{G_1 ... G_{n-1}} \] (2.1)

The sensitivity of a receiver, or the minimum power of a detectable signal, is then
defined as:

\[
S_{\text{sens,rx}} = kT + 10 \log(B) + \text{SNR}_{\text{min}} + NF_T \] (2.2)
where $kT$ represents the power spectral density of thermal noise, $B$ the noise bandwidth, $SNR_{min}$ the minimum required by the ADC to supply a sufficient BER, and $NF_T$ the total noise figure of the receiver chain.

As sensitivity and hence the weakest desired detectable signal demand minimum SNR and NF requirements, gain cannot be driven to be arbitrarily large as we are hit with the receiver blocks' degraded linearity performance and the ADC limitations (dynamic range – or maximum acceptable signal). Therefore, unlike what is expected, we note an increase in BER with a very large received signal strength – much like when the received signal is weak. Therefore, radio systems are not only bound on the lower end by the noise floor but also on the upper end by the inherent nonlinearities of the circuits.

System blocks are non-ideal and therefore do not perform always linearly – keeping in mind that some blocks’ functionality is based on non-linearity (mixers for example). However, if we look at receivers, nonlinearities bring rise to gain compression, intermodulation distortion, desensitization, and cross-modulation. These undesired effects could manifest when there is enough signal power either in the desired signal itself or possibly in a close-by interferer (or blocker). The 1dB compression point, $P_{1dB}$, represents the input power that experiences a 1dB decrease from the expected linear gain. When two closely spaced tones enter a non-linear system – say a desired signal and an interferer or possibly two interferers close to the desired band – mixing is bound to happen and intermodulation products appear in the spectrum, as depicted in Fig. 2.11. Second order intermodulation products fall far from the band of interest, at the sum and difference frequencies. While the former can be easily filtered, the latter is problematic in the case
of zero-IF receivers as it falls close to dc. Therefore, second order intermodulation is problematic for mixers. Third order intermodulations are the hardest to get rid of as they fall extremely close to the two tones and within the band of interest, in case of an LNA for example. Third order intermodulation is tested by applying two closely-spaced and equal power tones. The intercept point at which the third order intermodulations, that increase cubically with input power, equal the main tones (that increase linearly with input power) is called \( IP3 \), and consequently the input power at that point is called \( IIP3 \), as shown in Fig. 2.12. As with the total noise factor, the cascade of receiver blocks results in a total \( IIP3 \) measure given by:

\[
\frac{1}{IIP3_i} = \frac{1}{IIP3_1} + \frac{G_1}{IIP3_2} + \frac{G_1G_2}{IIP3_3} + \ldots
\]

(2.3)

![Fig. 2.11 Two-tone intermodulation spectrum](image-url)
The higher the gain at the first blocks the more stringent the linearity requirements for the preceding blocks. Additionally, care should be taken to never exceed the ADC dynamic range. Therefore, the receiver chain should be able to 1) provide a large enough gain to amplify weak signals to meet a minimum SNR by the ADC and 2) make sure that a strong received signal is not amplified beyond the following blocks' input range or that of the ADC. These two requirements might not result in a single converging link budget; one way to tackle this is by using a dual gain frontend with either the LNA/PA or mixer having two gain modes: A high gain mode to satisfy the sensitivity and noise requirements and a low gain mode to satisfy linearity constraints. Moreover, Automatic Gain Control (AGC) is also employed to cover a range of input powers to satisfy dynamic range requirements, or more commonly the spurious free dynamic range.
(SFDR). The SFDR is defined as the maximum input level where intermodulations do not exceed the noise floor in relation to the minimum detectable input power.

A careful link budget analysis is therefore necessary to realize a system that offers conformance to the standard’s specification while allotting the required performance specifications for the individual building blocks and components.

2.2.2 Component Metrics

The RF blocks that constitute the receiver or transmitter chains provide the required signal processing for establishing communication. Each of these blocks performs a certain task and suffers from a number of impairments limiting its effectiveness and performance. Receiver chains are more challenging, as they have to deal with the reception and demodulation of very weak signals. In this section, we briefly highlight the metrics and impairments of each of the blocks in the RF chain.

2.2.2.1 Signal Amplifiers: LNA and PA

The Low Noise Amplifier, as its name suggests, has to provide sufficient gain at the minimum amount of added noise. Hence its gain, $G$, and noise figure, $NF$, are of primary importance. Also, as this block interfaces to the antenna through a T/R switch or duplexer it has to present proper impedance matching for maximum power transfer of the weak received signal. The PA’s metrics are similar to those of an LNA but are more steered to providing sufficient power, good efficiency, and matching to the antenna.
Both LNA and PAs have to provide superior linearity. Designers strive to obtain the highest possible compression points and IIP3s to reduce the emergence of harmonic and intermodulation products. These spectral elements, if not treated appropriately with filtering or otherwise, can interfere with other frequency bands off-chip or on-chip. Off-chip, the PA should prevent spectral regrowth to adjacent channels that can be occupied by other users. Standards usually specify a certain transmit mask that products must adhere to. On-chip, the LNA’s output might contain third order intermodulation products as well as the main tone, with both delivered to the mixer for downconversion. Fig. 2.13 shows an ideal and non-ideal amplification, with the latter depicted as intermodulation distortion due to out-of-band interferers.

![LNA/PA Diagram](image)

Fig. 2.13 Ideal and nonlinear amplifier effects
Typically, one-tone and two-tone tests are performed on LNAs to quantify their gain, noise, and linearity parameters. Power detectors can be used at the inputs and outputs of these amplifiers to measure their signals and their contents and extract the important metrics.

### 2.2.2.2 Mixer

As described previously, the mixer is a three-port device with two inputs and one output. The signal of interest comes in from one port and exits the other. A local oscillator provides the other input port with a tone for frequency translation. The mixer ideally performs a sum and difference conversion in the frequency domain that is coupled with a gain (active mixers) or loss (passive mixers). Therefore, an important metric is the conversion gain as well as linearity (second order intermodulation is critical here) and NF. Also, the mixer’s conversion gain is a function of the input LO power and the two must be carefully designed in conjunction, especially in IQ-type systems.

Non-ideal isolation between the input ports results in self-mixing, feedthrough, and in-band intermodulation products (Fig. 2.14). A powerful LO can seep into the second input and mix with itself – its downconversion will become an undesirable dc offset. On the other hand, low isolation can also result in LO power to appear at the output close or coinciding with an upconverted signal in a transmitter.
2.2.2.3 Local Oscillator

The local oscillator is usually implemented as a frequency synthesizer, generating precisely controlled frequencies based on phase locking concepts. A Phase Locked Loop (PLL) uses feedback techniques to lock a high-frequency free running oscillator to a clean and steady low-frequency source.

PLLs are complex systems comprised of multiple types of circuits, spanning the digital, analog, and RF domains. In general, the inputs to the PLL are a clean low-frequency reference frequency \( F_{REF} \), usually supplied by a crystal, and a multiplication factor \( N \), a division value in reality) and the output is a stabilized high frequency
multiple of the reference \((N \times F_{REF})\). Fig. 2.15 shows a block diagram of a PLL. The negative feedback works on reducing the error between the divided voltage controlled oscillator (VCO) frequency and a stable crystal reference. The comparison block is the phase frequency detector (PFD) which outputs pulse-width-modulated error signals signifying the difference between the two inputs. The error/correction pulses, UP and DN, indicate a lead or lag between these inputs. The PWM signals are supplied to a charge pump (CP) that injects or removes a proportional amount of charges (current) into a smoothing filter. The latter converts the charge input to a voltage that eventually modulates the VCO.

PLLs as frequency synthesizers need to provide programmable frequency shifts to accommodate the different channel requirements dictated by the communication standards. Important synthesizer metrics are the center frequency, lock range, settling time, and phase error. The center frequency needs to be free of frequency and phase
offsets to ensure proper mixing and signal transmission/reception. The synthesizer has to be able to lock to all the frequencies of relevance and when instructed to change frequencies, do so in reasonable time (settling time).

A real LO signal is in reality not a clean sinusoid and does not resemble a single tone in the frequency domain, due to phase noise. Phase noise is considered one of the most important parameters in a frequency synthesizer. It is the frequency domain representation of jitter and manifests itself in frequency deviations from the ideal frequency (Fig. 2.16). Single-side-band (SSB) phase noise is measured in dBc/Hz and gives the noise power (normalized to 1 Hz bandwidth with respect to the carrier) at given frequency offsets from the carrier. Phase noise is also integrated over a frequency band to provide a single value, either expressed in dBc or in degrees as phase error. Root-mean-squared phase error can then be easily transformed to timing jitter.

Each component of the PLL contributes in one way or another to the phase noise of the output signal: some blocks contribute noise close to the synthesized signal while others’ noises dominate at higher offsets. Also, apart from random noise, high-powered discrete noise tones can appear in the output, called spurs. A common spur in PLLs is the reference spur which appears at $F_{REF}$ offsets from the center tone. It is mainly caused by the PLLs’ periodic correction of some of its blocks’ non-idealities (loop filter leakage and charge pump current mismatch).
Since the PLL is used as a local oscillator to perform frequency up- or down-conversion, the non-ideal tone with “skirts” endangers a correct transmission or reception of a signal. This can be seen in Fig. 2.17: the noisy local oscillator causes one of the weak RF tones to be distorted by the dominant phase noise of the stronger tone. Therefore, the various communication standards impose a mask on the phase noise spectrum of local oscillators.
2.2.2.4 IQ Modulators and Demodulators

IQ modulators and demodulators are simply a special construction of LOs and mixers. The principle of operation was described previously as a method for image rejection. Since digital data is encoded and decoded with quadrature content on the digital level, it needs to be transmitted and received without the mirror images corrupting each other. Quadrature LO signals need to be generated and fed to two mixers, one on each path, as is shown in Fig. 2.18. It is important here that the matching between the $I$ and $Q$ channels be superior as minor mismatches will reduce the suppression of the mirror images and corrupt the intended symbols.
The important metrics here are amplitude and phase mismatch. The 0 and 90 degrees signals have to be of the same amplitude and in perfect quadrature. Also, differences in the two mixers’ gains contribute to the mismatch. If the amplitude and phases of the paths are not matched, the image rejection ratio, which represents how much of the undesired image has been suppressed, will be adversely affected. The IRR as a function of amplitude and phase mismatch is given by:

$$\text{IRR}_{\text{dB}} = 10 \log \left[ \frac{\alpha^2 - 2\alpha \cos \phi + 1}{\alpha^2 + 2\alpha \cos \phi + 1} \right]$$  \hspace{1cm} (2.4)$$

where $\alpha$ is the amplitude imbalance (expressed as a ratio) and $\phi$ is the angle mismatch from perfect quadrature between the two paths. A value greater than 60dB is often desired but slight imbalances cause a disproportionate drop in that value. Fig. 2.19 shows how sensitive the IRR is to small fluctuations in the amplitude and phase matching.
2.3 Integrated Radio and System-on-Chip Testing

The previously discussed system and block metrics need to be verified to ensure proper functionality. With integrated RF SoCs, both digital and RF blocks coexist on chip and testing becomes more involved. On one hand, digital testing has reached a mature stage and its techniques have adopted well-known and widely applicable fault models [23]. On the other hand, testing for RF and analog, in general, still relies on checking for conformity to a set of design specifications. However, observing these performance parameters is quite a challenge as RF systems, especially highly-integrated transceivers, offer little accessibility to individual blocks. Limited accessibility limits the detectability of faults and restricts observability. This is especially true for the traditional RF testing
methods where powerful mixed-signal Automatic Test Equipment (ATE) and benchtop laboratory setups – such as rack-and-stack – require a sizeable investment in terms of actual cost to obtain and maintain, difficulty of interfacing to the internal nodes, and long test times.

A promising technique borrowed from digital design and only recently applied to RF is Built-in-Self-Test, or BiST. On-chip BiST provides an opportunity to control inputs and observe outputs of individual blocks as well as signal chains. This is enabled by the insertion of simple measurement circuits at critical internal nodes. These sensors can then provide readings indicative of signal properties of interest at inputs, outputs, or even the insides of individual on-chip blocks. Moreover, an extension to mere sensing is also the inclusion of on-chip test stimulus generators for a truly internal test cycle. BiST techniques can therefore be cost-effective alternatives to, albeit not necessarily more accurate than, traditional testing equipment for complex integrated RF SoC.

Although neither ATE, Rack-and-Stack nor BiST techniques are used exclusively, only the latter promises beyond-post-production test portability. This is a very important enabling technique to ensure the viability of microchips with an extended set of operating conditions including process, voltage, and temperature changes (PVT). This gains more importance when we discuss self-calibration and self-healing techniques, for which BiST is a required precursor.

Having successful and efficient on-chip testing paradigms not only increases the probability of detecting faulty blocks but also allows for post-silicon quality measures, of which on-the-fly tuning promises to offer robustness enhancement for heterogeneous
systems such as RF SoCs. It has become quite evident with the increased integration that manufacturing testing is not enough for validating a complex system. In-field and operational variability needs to be taken into account; and to lump all these test cases at the production testing stage is a daunting task. Therefore, there is a need for increasing detectability and allowing measurement paradigms to be integrated into the circuitry rather than being outside.

2.3.1 Built-in-Test Techniques

Migrating the testing on-chip, different techniques have been proposed and used, either stand-alone or as enhancement to traditional tests. Several low-cost testing methods exist such as loopback testing, alternate testing, and digitally-assisted testing. These methods are not mutually exclusive and can be used together to achieve the desired BiST functionality.

Loopback testing is among the most well known methods, especially for RF systems employing both a transmitter and receiver. Its advantages include low cost, extremely low hardware overhead (if any), and simple (usually a single-metric) testing. Cost and hardware overhead are very low because of the high levels of component re-use. The digital baseband and RF sections of the system are tested together by forming a loop from the transmitter to the receiver. A digital bit-stream is generated in the transmitter’s baseband, transformed and upconverted to RF, and then coupled to the receiver [24]. The coupling circuitry is usually the only additional overhead. This coupling can be achieved internally (on-chip) or externally (on-board). However, depending on the type of
duplexing the system uses, the hardware needed might slightly change. For example, at
the minimum, a simple controllable attenuator can be used between the transmitter’s pre-
amp and the receiver’s LNA in a zero-IF TDD system (Fig. 2.20). In contrast, systems
whose transmitters and receivers have a frequency offset, such as in FDD systems, proper
frequency translation is required prior to coupling. Narrowband systems might be able to
perform the translation in the transmitter baseband; broadband systems would need a
dedicated mixer with the attenuator [25][26].

Loopback testing often offers a single-metric measurement, usually the BER (or
EVM). BER measurements allow for effective fault detection as they can be used to trace
the degradation due to noise and gain impairments in the system. While this is an
attractive feature for pass/fail testing, a major downside to loopback testing is fault-
masking. Since testing is done end-to-end with the entire system treated as a single block,
there is no information on which component is causing failure in the system. This makes
fault-localization very difficult especially in complex RF and analog chains. Some
methods do exist to alleviate these issues, such as path sensitization and internal node
monitoring. In the first, specially crafted test pattern can be transmitted, looped back, and
upon reception analyzed for distinct signatures. Faults, if they exist, can be attributed to
specific blocks based on the received signal’s signature. This technique, however,
requires some behavioral modeling effort to quantify appropriate test patterns and their
responses. Another method is to augment fault observability by actually monitoring
internal nodes in the system, at the expense of additional hardware overhead. On-chip
sensors can be inserted to provide more signal information along the loopback chain [27].
Alternate testing is another on-chip low cost testing method that is geared towards characterizing individual components rather than end-to-end specification checking. Alternate tests do not attempt to sense or measure a certain circuit parameter directly but opt to get a reading that can be explained by a set of circuit parameters. Therefore, it presents an attractive alternative for decreasing test time with the possibility of extracting multiple circuit parameters in a single test, and predicting the specifications accordingly. In essence, alternate testing is a correlation testing methodology in which several spaces are mapped to each other: parameter, signature, and specification spaces (Fig. 2.21) [28]. Parameter spaces can be constructed from known circuit and process variations. Then suitable test stimuli are created to expose appropriate signatures in which several distinct circuit parameters are distinguishable. The response of the circuit to the test stimulus constructs the signature space. Both the parameter and corresponding signature spaces can be mapped to a specification space. The acceptance region (pass/fail zones) needs to be carefully determined and mapped. This may require effort on the side of modeling and
statistical simulations (Monte Carlo) or even actual measurements on test silicon runs with enough samples [29].

Fig. 2.21 Alternate testing: parameter, signature, and specification spaces and their mapping [29]

A main objective of alternate testing is the use of inexpensive test generation and signature detection. The alternate testing premise can be contrasted with the previously discussed path sensitization in loopback testing: Path sensitization is applied to systems with many cascaded blocks in order to enable extraction of a specific component parameter whereas alternate testing predicts multiple block parameters from tests on a single component. A downside of alternate testing is the need for solid mappings between the different spaces – process shifts that alter the statistics and correlations require retraining of the models and the corresponding spaces to retain the correct acceptance zones [29].

A testing approach emerging with the proliferation of complex transceiver chips is digitally-assisted testing. This testing paradigm builds on the design of digitally-assisted
circuits (Fig. 2.22). A number of transceiver blocks now comprise not only of analog but also digital parts with some of the analog tasks migrating to the digital domain. These circuits’ digital parts are used to internally monitor performance and tune the analog parts according to predetermined optimizations and calibrations, all in a closed loop manner.

Fig. 2.22  Digitally-assisted analog/RF circuit

With the built-in capability to digitally monitor and tune the analog performance, specifications can therefore be extracted by observing or reading the digital state. For example, many analog biasing voltages are supplied by DACs in an effort to overcome PVT variations with some degree of programmability; therefore, a reading of the digital word in the DAC can be indicative of the analog block’s performance. A more attractive aspect of digitally-assisted circuitry in general is the ability to embed tuning programmability into SoCs at minimal cost.
What comes to transpire then is the notion and possibility of combining the above on-chip testing techniques to allow more efficient built-in-self-test coupled with enhanced calibration capabilities. This is the topic of the upcoming chapters.

2.4 Summary

This chapter described transceiver architectures including a system and component overview. The most relevant metrics were presented and a brief discussion on the component non-idealities and their effects on proper system functionality. Built-in testing techniques are also described. The material presented in this chapter form the backbone of the suggested RF built-in-self-test and digital self-calibration to be discussed in the later chapters.
Chapter 3 | Efficient Testing for RF SoCs

The implementation of self-test, as discussed in this Manuscript, is not geared exclusively for production testing but also for post-production. Post-production, or more specifically, post-deployment testing will provide the platform for the implementation of on-chip adaptive calibration. The goal then is not to pass production testing only but to ensure that a product does not operate at the edges of the pass zone, but always comfortably in its optimal region. This is not only to satisfy silicon yield but also to ease the ever increasing complexity of the design process. Designing in nanometer CMOS is becoming a daunting task with the increase in numbers of corners, heralded by process variability and volatility with temperature and power. Faced with the non-optimality of over-design, circuit designers are looking into implementing more innovative techniques – those of assisted operation, therefore boosting performance of the otherwise non-optimal circuit with a robustness enhancer. The concept of tunable RF circuits, and what techniques are best suited to augment their capabilities in SoCs will be discussed in the next chapter. The challenge then becomes how to tune these circuits optimally, over corners, and over time, meaning while operating in a consumer setting. For optimum tuning, the operating condition of these circuits needs to be assessed, beyond the lab or
fab. This is only manageable by putting measurement capabilities into the system, hence self-test.

In this chapter, we lay down the requirements for RF SoC BiST and propose a setup that makes maximal use and re-use of available resources, with little additional circuitry. Then, we introduce a measurement circuit that balances the often conflicting requirements of efficient and accurate measurements, insertion non-invasiveness, wide dynamic range, and broadband operation. The proposed circuit is a modification on RF amplitude/power detectors whose design, characteristics, and three different implementations are described.

3.1 On-Chip Test Migration and Portability

To truly migrate the testing functionality on-chip and enable test setup portability, an efficient self-test paradigm for transceivers and radio SoCs should include the ability to complete the test loop internally. Also, the capability to test one or more blocks in the chain, as well as whole system characterization, should be addressed. The applicable tests and the controllability of the test signals are also of concern, especially that the test signals under consideration here are at radio frequencies and beyond.

Therefore, we opt for an architecture that inherits and includes desirable features from each of the built-in-testing techniques described in the previous chapter. Loopback testing’s emphasis on system-level testing (such as BER and EVM) needs to get decoupled to enable specification-based direct and alternate testing of individual blocks. Moreover, as these blocks move to digitally-assisted designs, the luxury of tuning them
and extracting more performance information from them becomes quite apparent for characterization purposes and self-calibration possibilities. Here is where we draw the line between regular built-in test (BiT) and built-in-self-test (BiST). The latter’s self-sufficiency and non-reliance on any off-chip external test equipment is something only possible with heterogeneous single-chip systems. Therefore we need to highlight the following requirements for an on-chip on-the-fly test and calibration setup:

1) *Test signal generation:* Depending on the type of metric sought, the testing technique will differ. The system should have a somewhat flexible test pattern and signal generation. Given the application in the context of a SoC, several resources can be used to generate such signals. To increase resource re-use and reduce overhead, the presence of a flexible DSP in the digital backend can be leveraged to customize test signals, in line with digital signal synthesis [30]. The test signals can be representative of symbols, such as patterns used for BER and EVM tests, or the digital equivalents of analog tones, for use in component level specification testing. The latter setup combines the DSP with the interfacing transmitter DAC and analog baseband to generate single or multi-tone test sources, at baseband frequencies. A sufficiently wideband analog baseband is useful in such cases to create a wider range of synthesizable signals. These lower frequency signals can then be appropriately modulated to RF and be routed internally.

2) *Test signal routing:* The end-to-end test-bench is analogous to the loopback testing technique whereby a transmitter-receiver coupling, or signal routing, is achieved by switches, a programmable test attenuator, and possibly a mixer for transceivers
with frequency offsets (e.g. FDD). Insertion of these loopback blocks should not affect the regular operation of the transceiver, and need to be designed accordingly [30][32]. These blocks should match to the output of the PA and input of the LNA, taking in a relatively large and powerful signal and outputting a controllable signal. Also, it is preferable that they demonstrate superior linearity – putting passive implementations at a slight performance advantage but at higher area cost. Resistive ladder attenuators as well as MOS-based implementations offer good linearity, matching, and range of attenuation [1]. We note here that continuous attenuation is not really necessary and a finite and discrete set of attenuation levels suffices. A passive mixer in the loopback element is preferred as it has high linearity and inherent attenuation (loss) – with the possibility of having a programmable conversion gain. Both the attenuator and mixer can be programmed to vary their gain and therefore create discrete sweep scenarios. The range of attenuation should encompass all testing needs. For example, the lower limit of attenuation should still allow for tests that require large signals, such as compression tests, to be performed between transmitter and receiver. Therefore, it is very important to analyze these critical loopback elements [33].

3) **Internal node accessibility:** Apart from the loopback element that establishes the path between the transmitter and receiver and enables variable attenuation levels, RF components in the middle of the chains cannot be directly accessed for characterization and individual testing. Bypassing techniques can be used to route signals to a specific block while also turning off bypassed circuits for better signal
integrity. Low insertion loss switches are needed with very high isolation in the off state to ensure that signals pass through with minimal degradation. To decrease the insertion loss, bigger switches are needed however this increases their capacitive contributions at high frequencies. The size of the switches should be taken into consideration when designing the rest of the circuits [34].

4) *Internal node visibility:* One of the downsides of loopback testing is the lack of information on the internal nodes of the end-to-end setup. This has been alleviated by the inclusion of small on-chip RF detectors that attach to these nodes and provide an easily readable measure of the RF signal. A common output of these detectors is a dc value corresponding to either peak amplitude or RMS power. We will leave the more extensive description to the next section that presents existing RF detectors and proposes a more suitable implementation for RF SoC BiST. However, we mention here a few notes pertaining to this block. RF detectors, much like the test attenuator and switches, are additional circuits and as such their presence needs to be as non-invasive to the system as possible. This places certain requirements on their design. Multiple detectors can be placed along the transmitter and receiver chains, with their outputs forming a low frequency (mostly dc) bus that can be easily digitized for analysis.

5) *Test result analysis:* The powerful digital backend can also be used to perform test results analysis. Test results can be derived from two sources: the primary digital lane and the auxiliary dc lane. The digital lane is simply the digitized *I* and *Q* channels on the receiver. This is in essence the regular return test path for a
loopback BiT where two channels interface to the digital processor through their respective ADCs – which is also the case under normal transceiver operation. On the other hand, the auxiliary dc lane is a test-only path that holds the test data from the RF detectors. These dc values can be digitized using the system ADCs and their readings used to extract performance parameters.

6) **Overhead:** An important recurring point in the previous discussion is the stress on lowering the overhead – in area, power, and test time – of the new embedded capabilities. Regarding area, a suitable architecture will make maximal use of existing hardware and only require a few additional circuits (attenuator, switches, offset mixer, and detectors) – normally well below 10% of the total area [35]. The major benefit here is resource re-use, and more importantly the powerful digital hardware which when coupled with a malleable software layer (e.g. algorithms) can offer high levels of flexibility and customization. As testing is only performed intermittently, power and time overhead can be easily optimized. When not testing, the additional circuitry can be turned off, e.g. power gating. Moreover, testing can be scheduled by the system during down time.

The next section presents a complete transceiver with the required modifications for enabling RF BiST and BiSC.

### 3.2 A BiST-ready RF SoC

Fig. 3.1 shows the modification of the transceiver architecture previously presented in Chapter 2. The changes are done to satisfy the requirements set forth in the previous
RF signals can be created in the baseband of the transmitter and upconverted by the mixers. A router-type loopback element interfaces the transmitter to the receiver. Switches accomplish routing of the test signals and their attenuation is further controlled by the programmable attenuator. Bypassing the front elements of the transceiver also deactivates them so that they do not affect the signals being routed around them.

Fig. 3.1 A BiST-ready RF SoC

The system shown here is a TDD system where a single LO is used for the transmitter and receiver. In the case of an FDD system with two LO blocks, the frequency offset between the receiver and transmitter can be synthesized in the transmitter baseband or appended in a loopback mixer. All blocks in the system, even the loopback elements, can be monitored by the RF detectors that are scattered around the two chains. The presence of this small detector eliminates fault masking and enables the testing and monitoring of the test generation circuitry before signals are routed for
testing. This ensures test signal integrity and enables signal tracking along the chain. The multiple detectors’ outputs form two dc busses going to the receiver ADCs. A main multiplexer toggles the input of the ADCs between normal and test modes, sending the demodulated quadrature signals in the first case and detectors outputs in the second. Successive detectors are assigned to separate ADCs to allow for concurrent input and output readings of a single block. In the case that the required detectors outputs are assigned to the same ADC, then readings can be time interleaved.

Internal to the digital part is the DSP responsible for the testing algorithms. The algorithms specify the test signal creation and control the test circuitry, i.e. switches, attenuator, and detectors. They also perform the analysis to extract the blocks’ performance parameters. Table 3.1 highlights the important parameters to test for in an RF SoC. Several of these parameters can be extracted using power and amplitude measurements while others require spectral measurements. The general one-tone and two-tone tests can be generated in baseband for parameter extraction. Compression and intermodulation points can be measured by successive sweeps of these tests. Other parameters can be indirectly deduced from alternate test result processing. Possible measurements in this system will be highlighted later in this chapter.

<table>
<thead>
<tr>
<th></th>
<th>LNA</th>
<th>Mixer</th>
<th>PLL (LO)</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gain</strong></td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Output power</strong></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>Linearity</strong></td>
<td>●</td>
<td>●</td>
<td></td>
<td>●</td>
</tr>
<tr>
<td><strong>Input Match</strong></td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>I/Q Match</strong></td>
<td></td>
<td></td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td><strong>Noise</strong></td>
<td>●</td>
<td>●</td>
<td></td>
<td>●</td>
</tr>
<tr>
<td><strong>Phase Noise</strong></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

Table 3.1  Important transceiver parameters to measure
A more intriguing capability now arises with circuit self-awareness. Based on the results of the tests, the transceiver blocks can be tuned to optimize their performance. As the circuit implementations of the RF blocks makes use of the digitally-assisted methodology, calibration DSP algorithms can make sense of test results to change the state of the block under calibration. BiSC will be discussed in Chapter 5.

Earlier implementations of loopback with increased observability through RF detectors still relied on outside equipment to properly analyze the output signatures of the detectors [35][36][40]. True BiST should only use internal components and therefore maximally benefit from the ADC and DSP at the back of the high-frequency transceiver. The measurement, or reading, accuracy is then limited by the internal ADC resolution and the RF detectors’ sensitivity. Higher resolution ADCs will enable better discrimination between dc outputs and high sensitivity detectors will offer better distinction between slight changes in RF features. The ADC resolution is primarily set by the system requirements, with ranges in the order of 10- to 12-bits in state-of-the-art ADCs. Therefore, to improve testing accuracy, the RF detector’s sensitivity should be increased.

3.3 RF Amplitude Detectors for RF BiST

This section describes the sensor to be used in the on-chip BiST and BiSC. The RF amplitude detector presented here is capable of broadband operation with wide dynamic range and high sensitivity. The requirements for such a detector for successful embedding
into a self-test scheme are listed along with previous detectors in literature and their shortcomings. The most prevalent types of detectors convert high-frequency signals – either sensitive to amplitude or power of a signal – to a corresponding dc voltage. This dc voltage is a digitally-friendly mapping of the RF signal which upon digitization will represent a reading of the signal state. Following in this section, the proposed amplitude detector is presented along with its analysis, design, and circuit implementations.

3.3.1 Detector Requirements

For parameter extraction of the various blocks in the transceiver, multiple detectors need to be inserted along the RF chains. Since the signal levels and properties differ between these blocks, a suitable detector needs to be designed to ensure accurate measurements chain-wide. Moreover, much like the additional test circuitry, the detector should be as non-invasive as possible so as not to affect the normal operation of the block and signals it is monitoring. The design of an appropriate detector for RF SoCs should meet a number of requirements [40]:

1) *Small area and low power:* These are major requirements, as multiple detectors are needed in the loopback chains. A detector should be only a fraction of the main block it is monitoring which limits the type of circuits and device sizes that can be implemented. Power is less of an issue if the detector can be turned off when the system is not in test mode.

2) *Non-invasiveness:* The detector has to connect to the signal path but should be designed not to load it. Therefore, it should be transparent to the system.
Designing the detector with a high input impedance can ensure minimal loading, which is critical in the impedance matched blocks like the LNA and PA.

3) *Wide dynamic range:* The dynamic range represents the range of amplitudes that can be sensed. In an effort to have a single detector implementation for the entire system, the designed circuit should withstand widely varying amplitude signals while outputting a correct dc value.

4) *Broadband operation:* The detector’s ability to cover a wide range of frequencies will enable its use in multi-standard RF SoCs.

5) *Accurate and sensitive response:* A stable and accurate high-frequency-to-dc conversion will allow for very fine detection of small amplitude changes. A stable detector response ensures that PVT variations will not affect the measurement while a high conversion gain from RF to dc will decrease the minimum detectable amplitude change, given by

\[
\min(\Delta \text{amp}) = \frac{V_{\text{fullscale}}}{2^{\text{ADCbits}} A_{\text{rf-dc}}} \quad (3.1)
\]

where \(V_{\text{fullscale}}\) is the ADC fullscale, \(\text{ADCbits}\) the number of bits of the ADC and \(A_{\text{rf-dc}}\) the RF-to-dc conversion gain.

The above requirements place guidelines for the design of a suitable detector for inclusion in an RF BiST scheme.
### 3.3.2 Detector Architectures

CMOS on-chip amplitude detectors have recently gained importance for Built-in-Test applications to enable internal probing of RF transceivers. Several implementations have been published in literature [37]-[46]. The basic premise is the use of the non-linear properties of MOS transistors to convert the high frequency input signal to a low frequency (and dc) current, which in turn creates a voltage over a load. Single-ended and differential implementations are possible depending on the type of the circuit being monitored. Table 3.2 lists some properties of the CMOS RF and mm-wave detectors in literature and Fig. 3.2 plots their characteristic conversions.

<table>
<thead>
<tr>
<th>Technology</th>
<th>[37]</th>
<th>[38]</th>
<th>[40]</th>
<th>[41]</th>
<th>[43]</th>
<th>[45]</th>
<th>[46]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area (mm²)</td>
<td>-</td>
<td>0.0043</td>
<td>0.0016</td>
<td>0.06</td>
<td>0.03</td>
<td>-</td>
<td>0.045</td>
</tr>
<tr>
<td>Frequency (GHz)</td>
<td>1 – 5</td>
<td>- (1)*</td>
<td>0.1 – 20</td>
<td>- (5.2)*</td>
<td>0.9 – 2.4</td>
<td>- (60)*</td>
<td>- (70)*</td>
</tr>
<tr>
<td>Dynamic Range</td>
<td>0.05 –</td>
<td>0.05 –</td>
<td>0.05 –</td>
<td>0.05 –</td>
<td>0.02 –</td>
<td>0.03 –</td>
<td>0.03 –</td>
</tr>
<tr>
<td>Conversion Gain</td>
<td>V/V</td>
<td>V/V</td>
<td>V/V</td>
<td>23mV/d</td>
<td>-50 mV/</td>
<td>+0.72</td>
<td>+1 V/V</td>
</tr>
<tr>
<td>Loading</td>
<td>7.6fF</td>
<td>21fF</td>
<td>12fF</td>
<td>-</td>
<td>13fF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Power (W)</td>
<td>3.6μ**</td>
<td>0.6m</td>
<td>10μ**</td>
<td>3.5m</td>
<td>8.6m</td>
<td>-</td>
<td>0.5m</td>
</tr>
</tbody>
</table>

* a range is not specified  ** only detector core, without biasing circuits

Table 3.2 RF detectors in literature
From Fig. 3.2, it can be seen that most detectors have a low conversion gain and some have a narrow amplitude detection range. An architecture that has both a high conversion gain and wide amplitude range is presented in the next section.

### 3.3.3 Proposed Detector Design

The proposed detector, shown in Fig. 3.3, is a simple single-stage implementation quite similar to some of the previously mentioned implementations [37][40][46]. However, its implementation includes modifications to increase the conversion gain and widen the dynamic range. The input of the detector is ac-coupled to allow for separate biasing to the gate of the input device. The output is taken out at the low-pass filter end of
the detector. The load is either an active device or simply a resistor – in the actual implementations discussed later either is used. The detector presents a capacitive load to the path it is monitoring; hence, its input capacitor and device need to be sized to exhibit a large impedance at the frequencies of interest.

![Proposed detector architecture](image)

**Fig. 3.3** Proposed detector architecture

The input device is biased in the subthreshold region. Subthreshold conduction offers a number of benefits including a higher transconductance (i.e. more sensitive V-I transfer) and very low power consumption due to the much lower current (at the expense of slower response). The saturated drain current in the subthreshold region can be expressed as [47]

\[
I_n = I_{D0} \left( \frac{W}{L} \right)_n \exp \left( \frac{V_{GS}}{nV_T} \right)
\]

\[(3.2)\]
where $I_{D0}$ is a current constant independent of gate-to-source voltage, $(W/L)_n$ the aspect ratio of the NMOS transistor, $n$ a process dependent term related to depletion region characteristics, and $V_T$ the thermal voltage ($=kT/q$; 26mV at room temperature).

With a small sinusoidal input $V_a\cos(\omega t)$ superimposed on the gate bias $V_{bias}$, the following power series approximation of the current equation holds

$$I_n = I_{D0} \left( \frac{W}{L} \right) \frac{V_{bias} + V_a \cos(\omega t)}{V_T} = I_{B0} \exp \left( \frac{V_a \cos(\omega t)}{nV_T} \right)$$

$$= I_{B0} \left[ 1 + \frac{V_a}{nV_T} \cos(\omega t) + \frac{1}{2} \left( \frac{V_a}{nV_T} \right)^2 \cos^2(\omega t) \right]$$

$$= I_{B0} \left[ 1 + \frac{V_a}{2nV_T} + \frac{V_a}{nV_T} \cos(\omega t) + \frac{V_a}{2nV_T} \cos(2\omega t) \right]$$

(3.3)

where $I_{B0} = I_{D0}(W/L)_n \exp(V_{bias}/nV_T)$ is the dc-bias current of the transistor. The larger the amplitude, $V_a$, the larger the drained current and the further $V_d$ is pulled away from $V_{DD}$ causing the output node ($DC_{out}$), charged initially to $V_{DD}$ by $I_{bias}$, to discharge thus establishing a negative relation with respect to the RF signal amplitude. As the detector output is low-pass filtered by the RC load, it reacts to the frequency-independent (but temperature dependent) dc component of $I_n$ given by

$$I_{dc} = I_{B0} \left[ 1 + \left( \frac{V_a}{2nV_T} \right)^2 \right]$$

(3.4)

This discharging dc current then becomes a dc voltage at the load. Proper sizing of the load and the input device will enable a high RF-to-dc conversion gain; however, with a limited voltage supply, increasing the conversion gain limits the dynamic range. To counter that, the detector is operated beyond its regular mode. By simply changing the
gate bias of the input device while keeping the latter in the subthreshold region, the same detector characteristic can be extended to other amplitude ranges. This will essentially create extended operation modes with sub-ranged detection regions (Fig. 3.4). For higher amplitude signals, $V_{bias}$ is made smaller such that the NMOS transistors are turned on with larger RF signal amplitudes. A digitally-programmable voltage bias circuit can then be implemented to shift the detector characteristic to the amplitude range of interest. The voltage shifts, or modes, can be designed to create continuous or overlapping regions with specified offsets. Upper and lower limits on the output dc value can then be used to automatically shift to the previous or next mode, respectively. These limits can be chosen around the range where the conversion gains of all modes are identical. The combined response is capable of covering a wide range of amplitudes with a high conversion gain.

Fig. 3.4 Proposed detector characteristics
Knowing the state of the detector, a dc output can be mapped to its corresponding RF amplitude reading as

\[
V_{\text{meas}} = \frac{(dc_{\text{hi}} - dc_{\text{out}})}{|A_{\text{RF-dc}}|} + \text{offset}_x \tag{3.5}
\]

where \(V_{\text{meas}}\) is the extracted reading, \(dc_{\text{hi}}\) the set limit of the detector output at zero signal, \(dc_{\text{out}}\) the detector’s actual dc output, \(A_{\text{RF-dc}}\) the conversion gain, and \(\text{offset}_x\) the offset between modes.

Of importance also is the applicability of two-tone signals in the BiST routines. Therefore, the behavior of the detector should be described under such inputs. Considering two closely spaced sinusoidal inputs \(V_a\cos(\omega_1 t)\) and \(V_b\cos(\omega_2 t)\) superimposed on the gate, several intermodulation products arise. However, due to the low pass nature of the load most of the non-linear components are attenuated except for those at the low delta frequency of \(\omega_2 - \omega_1\) \((= \Delta F)\). In such case, the low frequency component of the discharging current \(I_n\) appearing at the output is then

\[
I_{n,\text{low}} = I_{B0} \left[ 1 + \left( \frac{V_a}{2nV_T} \right)^2 + \left( \frac{V_b}{2nV_T} \right)^2 + \left( \frac{V_aV_b}{(2nV_T)^2} \right)^2 + \frac{V_aV_b}{2(nV_T)^2} \cos(\Delta F t) \right] \tag{3.6}
\]

The output is then a low frequency oscillating signal that can be easily digitized by the ADC. The average dc value of the oscillating output represents the contribution of two tones. If \(V_a\) and \(V_b\) are equal, the average dc output of the two-tone is approximately equivalent to the dc output of a single \(\sqrt{2}V_a\) tone. That is the case for ideal two-tones, as is shown in Fig. 3.5. In a non-ideal case, the average dc output of the detector is also affected by the presence of other tonal elements, especially intermodulations of the third order that also appear close to the two tones. This forms the basic premise behind the
detector’s use for intermodulation distortion and linearity parameter extraction, to be discussed in Chapter 4.

![Diagram](image)

Fig. 3.5 One- and Two-tone transient responses with all tones having same amplitude

Also, the oscillating low frequency output can be used to verify the tone spacing as its oscillation frequency is $\Delta F$. Therefore, the detector’s output under a two-tone test scenario can be used as a measurement for amplitude and also a verification of the test-signal’s frequency spacing. The first measurement is useful for testing the signals at the inputs and outputs of a circuit-under-test (CUT) whereas the second can be used by the system as integrity check for the test-generation circuitry.

The use of an RF detector for test is mainly reliant on its accurate prediction of signal amplitudes. However, in calibration routines (BiSC) that depend on signal amplitude comparisons between calibration steps, only relative accuracy is necessary - meaning that a guarantee that the detector response is monotonic will suffice. Therefore for BiST applications, a highly accurate detector is required. To the first degree, it can be seen that the discharging current is a function of temperature and as such will create
slight shifts in the characteristic response. Second, in smaller CMOS processes, variations in the devices severely affect the operation of the detector, for example the threshold voltage variations (see Table 1.2 in Chapter 2). Then, the detector would require calibration to characterize and stabilize its RF-to-dc conversion. For a fully embedded and standalone BiST approach, that calibration should be performed on-chip and not through separate testing.

In the next section, several implementations of the detector are described in various process nodes and for different high-frequency ranges. These implementations will showcase differential and single-ended designs aimed at both the microwave (RF) and millimeter wave bands. Also, a design that overcomes variations by self-adjusting and aligning its response is presented.

### 3.3.4 Implementations for RF and mm-wave BiST

In this section we present a number of implementations of the proposed detector, namely a detector covering the 0.5 – 9 GHz range, 10 – 30 GHz range, and the 60 GHz band.

#### 3.3.4.1 Microwave Implementation in 180nm CMOS

Although an old technology by today’s standards, 180nm CMOS is still used for a number of RF applications as it is relatively inexpensive and able to cover a number of
widely used standards such as Bluetooth, WiFi, DECT, and even any upcoming applications in TV whitespaces (e.g. IEEE 802.22 Wireless Regional Area Network).

The detector circuit is built in 180nm CMOS technology from TSMC [48]. Fig. 3.6 shows the detector’s pseudo-differential core with an active pMOS load and output low pass filter. The input nMOS devices are chosen to be RF transistors and the input capacitors as MIM (Metal-Insulator-Metal).

Fig. 3.6 Pseudo-differential RF amplitude detector core

The input stage devices are biased through large resistors by a separate programmable biasing circuit. The biasing circuit, shown in Fig. 3.7, is a simple two-bit voltage divider providing four discrete biasing voltages. The entire detector (core + biasing) circuit runs from a 1.8V supply and consumes less than 400µA (~0.8mW) when in operation. To save on power, the detector’s pMOS load can be turned off effectively shutting down the core.
The characteristic response of the detector under a differential 2.5GHz RF input signal is shown in Fig. 3.8. The zero-RF dc output was set at 1.6V by design. It can be seen that the four modes of operation cover a continuous range from 0 to 0.7V amplitudes with each mode having a linear region between 1.6V on the upper end and 0.2V on the lower. Therefore, these two endpoints are used to construct the combined response as shown in Fig. 3.9, revealing a conversion gain of -10V/V. The detector also exhibits a very broadband range of frequencies, extending from 500 MHz to 9 GHz, over which the response holds constant. Moreover, the detector’s input impedance is larger than 8.5 kΩ over the frequencies of interest.
Fig. 3.8 Response of microwave amplitude detector at 2.5GHz

Fig. 3.9 Combined continuous linear response of the detector for input frequencies between 0.5 and 9 GHz
Of note here is the deviation of the dc response at very low signal amplitudes. It can be observed that the linear relationship does not hold at signal amplitudes below 100mV where it becomes a square relation, as depicted in Fig. 3.10. This discrepancy should be taken into account when processing the detector’s output. For example, when operating in the first mode and the detector’s dc output is greater than 1.35V, then the linear approximation of equation (5) does not completely hold and needs minor modification. The curve can be split into two at the dc output corresponding to 100mV RF input, \( dc_{sq} \), and the following can be used to determine the measured RF amplitude,

\[
V_{\text{meas}}_{\text{offset}} = \begin{cases} \frac{(dc_{hi} - dc_{out})}{100(dc_{hi} - dc_{sq})} & \text{when } dc_{out} \geq dc_{sq} \\ \frac{(dc_{sq} - dc_{out})}{|A_{RF-dc}|} + 0.1 & \text{when } dc_{out} < dc_{sq} \end{cases} \quad (3.7)
\]

Fig. 3.10 Square-law and linear regions of the detector response at low RF signal amplitudes
The frequency-independent response however is not immune to process and temperature changes. Fig. 3.11 shows the response curves for a number of possible variations in operating conditions. The deviation from the ideal case, which is used to build the baseline dc-to-amplitude mapping function, will then result in very large amplitude measurement errors. One simple and passive method to partly remedy this issue is to perform a one-point calibration by re-referencing the zero-RF dc output to the ideal value \(d_{\text{ch}}\), i.e. “zero-RF re-referencing”. For example, the zero-RF dc output at 60°C is 1.47V rather than the ideal 1.6V. Knowing that, the 1.47V can be virtually referenced as 1.6V in the DSP after the RF detector’s output is digitized by the ADC. Fig. 3.12 shows the measurement accuracy with and without zero-RF re-referencing for the first mode of the detector. It can be seen that measurement accuracy is increased by limiting the dc-to-amplitude mapping error to less than 10% at low amplitudes and less than 5% onwards. While this technique works on this CMOS node, it might not be as easy on processes that exhibit much larger variations. Therefore, a method that can stabilize the detector’s response with PVT is presented in Section 3.3.4.3.
Fig. 3.11  Deviations of the detector's response due to temperature and process variations

Fig. 3.12  Zero-RF re-referencing to increase measurement accuracy
3.3.4.2 Millimeter-wave Implementation in 90nm CMOS

A single-ended implementation of the detector is also designed targeting the mm-wave spectrum (60 GHz ISM band) using 90nm CMOS from IBM and powered by a 1.2V supply. This implementation has a mode-select programmable bias achieving 8 overlapping modes. The detector core and biasing are shown in Fig. 3.13. The mode offsets are set to 50mV. With a -9V/V conversion gain, the covered amplitude range is between 0 and 0.5V, as shown in Fig. 3.14. The upper and lower dc limits are set at 1V and 0.2V, respectively. Similar to the earlier implementation, the mm-wave detector response is to a large extent frequency-independent showing the same characteristic for the 55 GHz to 65 GHz range.

Fig. 3.13 mm-wave amplitude detector core and bias circuits
At extreme high frequencies such as the band under consideration, the capacitive load presented by the detector starts losing its high impedance. In lieu of input impedances in excess of 8-10kΩ in the RF implementation, this mm-wave detector exhibits impedances at an order of magnitude less, between 800-1000Ω. This still presents a small loading but should be accounted for in the design of the mm-wave system.

3.3.4.3 Self-adjusting detector implementation in 65nm CMOS

The one-point passive calibration method shown earlier to counter process and temperature mismatch does not track well with smaller CMOS processes where slight
shifts in operating conditions throw the characteristic curve way off. A more involved calibration method is thus required.

There are two pivot points that can be used to stabilize the detector’s characteristic curve. One is the zero-RF dc value, \( dc_{hi} \), and the other is at the other end of the curve, i.e. at the last mode’s \( dc_{lo} \). If these two points can be fixed then the characteristic curve will be held in place with PVT variations.

Fig. 3.15 shows the detector implementation, in 65nm CMOS from IBM, for the 10 to 30 GHz frequency range with in-built self-adjustment capabilities. The design makes use of two replica cores, zero-RF and max-RF replicas, placed in proximity to the main core to reduce variations. Also, all the cores’ input devices are sized at much larger than minimum length to have better matching between all copies, while keeping their input impedances relatively high (> 1kΩ at 30 GHz).

The first replica replaces the fixed programmable biasing of the previous implementations and works on fixing the upper dc output limit, \( dc_{hi} \), to 1V. Therefore, no RF signal is applied to that replica. Its output is compared to a fixed 1V reference by an operational amplifier (op-amp), which through feedback adjusts that replica’s gate voltage to force the appropriate bias point for \( dc_{hi} \). An analog subtractor, described in [49], is inserted in the feedback path to enable a programmable mode-select. The latter takes the op-amp output and generates two bias voltages that are at a programmed offset from each other. This enables the main core to have a lower bias voltage for its extended modes while the replica core sets the relative starting bias voltage.
The second point in need of stabilization is at the end of the characteristic curve. For that, another replica is included and is supplied the maximum RF amplitude – in this case, a rail-to-rail high-frequency oscillation. The max-RF replica is set at the last desired mode offset \( V_{\text{max}} \) also through an identical subtractor and its input is derived from a small inverter-based ring oscillator. The oscillator is built with a minimum sized three-inverter chain and buffered to the input of the max-RF core. The ring’s frequency of oscillation is slightly higher than 15 GHz; but since the detector is broadband and frequency-insensitive, deviations in that frequency do not distract from the replica’s functionality. The function of this replica is to adjust the loads of all cores such that its dc
output at maximum RF input and last mode of operation is exactly $dc_{lo}$. This is forced through an op-amp regulating the pMOS load and comparing the max-RF replica’s dc output to a 0.2V reference.

The result is shown in Fig. 3.16 with changes in temperature, process, and input frequency resulting in extremely minimal and indiscernible shifts in the primary and extended modes of operation. The bias offset ($V_{mode}$) was selected to provide 4 slightly overlapping modes – whereby the last offset (the one applied to the max-RF replica, $V_{max}$) determines how steep the conversion gain is. Monte Carlo statistical simulations are also performed to enable a more realistic characterization of this method since mismatches between the cores, main and replicas, will affect the detector’s actual curves. More than 500 Monte Carlo runs with process and mismatch variations over a temperature range of -30 to 90°C show a maximum detection error of 8% at the lowest end, quickly reducing to less than 2% at the extended modes (Fig. 3.17).

One feature of this configuration is the ability to adjust the conversion gain. If, for example, both the main and max-RF cores are operated in the primary mode, then a single mode covers the 0 to 0.6V RF amplitudes thereby decreasing the conversion gain. On the other end, if the max-RF core is operated at an even further offset than shown in Fig. 3.16, then the conversion gain increases.
Fig. 3.16 Self-adjusting RF detector modes under different temperature and process variations

Fig. 3.17 Maximum detection errors (Monte Carlo simulations) across the amplitude range
3.3.4.4 Comparison of implemented detectors

The specifications of the designed RF and mm-wave amplitude detectors are summarized in Table 3.3. These detectors will be used in the later chapters for BiST and BiSC. As stated previously, a highly accurate implementation is preferred in the case of testing but small performance shifts can be well tolerated when using the detector for calibration purposes, as will be shown in a later chapter.

<table>
<thead>
<tr>
<th>Implementation</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>0.18µm</td>
<td>90nm</td>
<td>65nm</td>
</tr>
<tr>
<td>Frequency Range</td>
<td>0.5 – 10 GHz</td>
<td>55 – 65 GHz</td>
<td>10 – 30 GHz</td>
</tr>
<tr>
<td>Dynamic Range</td>
<td>0.01 – 0.65V_{amp}</td>
<td>0.01 – 0.5V_{amp}</td>
<td>0.01 – 0.6V_{amp}</td>
</tr>
<tr>
<td>Conversion Gain</td>
<td>-10 V/V</td>
<td>-9 V/V</td>
<td>-5 V/V*</td>
</tr>
<tr>
<td>Loading</td>
<td>&gt;8.5kΩ</td>
<td>1kΩ</td>
<td>1kΩ</td>
</tr>
<tr>
<td>Area</td>
<td>0.08×0.08 mm²</td>
<td>-</td>
<td>0.15×0.10 mm²</td>
</tr>
<tr>
<td>Power</td>
<td>0.8mW (1.8V)</td>
<td>0.5mW (1.2V)</td>
<td>3mW (1.2V)</td>
</tr>
</tbody>
</table>

* can be adjusted

Table 3.3  Comparison of the implemented RF amplitude detectors

3.4 Summary

In this chapter, we discussed the requirements for migrating RF test to inside the chip therefore enabling testing even after the production stage. Using the various on-chip resources, a BiST-ready RF SoC is presented with the additional enabling circuitry. One of the essential blocks is the RF sensor, mostly implemented as a power or amplitude detector. We also highlighted the design requirements for RF detectors and showed a number of different implementations. We then proposed a similar design but with
different capabilities, enhanced for true on-chip testing in RF SoC. Three implementations in different CMOS technologies and covering the RF and mm-wave bands are discussed and their performances compared. In the next Chapter, we will put this detector to use in various testing schemes and routines.
Chapter 4 | RF Built-in-Self-Test

This chapter expands on the previous discussion on the BiST-ready RF SoC and the design of suitable RF sensors for true on-chip test. We place the RF amplitude detector designed in the previous chapter in the test loop to enable a number of important characteristic tests for RF and mm-wave blocks. We first describe the overall BiST routine that makes use of the digital core, transmitter circuits and loopback elements to generate test signals and route them accordingly to accomplish specification testing of most RF blocks through simple one- and two-tone tests. Then we demonstrate the effectiveness and viability of the RF detector in predicting and quantifying signal and circuit parameters through simulation examples in various frequency bands and process technologies.

4.1 Specification-based Tests using the RF Amplitude Detector

In this section, specification-based testing of various RF blocks is described with the use of the developed RF detectors. Signal amplitude can be indicative of various RF block parameters and as such its proper detection and extraction by means of an accurate detector enables. It was shown in the previous chapter that the detectors have an inverse
relationship to the RF amplitude, i.e. the dc output decreases with increased RF amplitude. The detector responses are used to map a dc value to a corresponding RF amplitude, in other words amplitude prediction or extraction.

Here we describe the methods by which the detector can be used to extract RF block performance parameters such as gain, compression point, intermodulation distortion, and quadrature mismatch. The BiST-ready SoC architecture presented in Chapter 3 is repeated here (Fig. 4.1) with node numbers for convenience and ease of referencing.

![BiST-ready SoC architecture with the important measurement points along the transceiver](image)

Fig. 4.1 BiST-ready SoC architecture with the important measurement points along the transceiver

Depending on the block being measured – the circuit under test (CUT) – some bypasses and attenuation might be required. The detectors can be also used to ensure that the appropriate test signal is being routed to the CUT. Care should be taken to provide the CUT with suitable test signals, for example ones that do not send it into compression, for
certain types of tests. Therefore, sufficient attenuation can be achieved by controlling the test-generation and test-routing circuitry. For testing the PA, the previous blocks should have tunable components such as the upconversion mixer or LO – with preference on the earlier. For the receiver chain, for example, the test attenuator (and offset mixer, if any) should be tuned accordingly. The tuning of these test-generation and test-routing elements can be monitored and controlled because of the presence of amplitude detectors at their nodes – rendering them like any other CUT.

The BiST routine starts with creating the required test signal in baseband and upconverting to RF. The amplitude of the test signal can be monitored at node (7), or node (8), and adjusted if a tunable mixer is implemented. Testing the test-generation circuitry increases confidence in the overall setup and test results. For testing the receiver, the loopback element is activated and depending on the type of test and signal level, the test signal can be routed from (7) or (8) to nodes in the receiver. Table 4.1 lists the various circuits under test and their testing setups while stating which nodes to connect and to monitor, and what blocks can be turned off.

<table>
<thead>
<tr>
<th>Circuit under Test</th>
<th>Loopback Connect</th>
<th>Disable</th>
<th>Nodes to Monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNA</td>
<td>(7) ⇒ (1)</td>
<td>PA</td>
<td>(1) &amp; (2)</td>
</tr>
<tr>
<td></td>
<td>(8) ⇒ (1)</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>No loopback</td>
<td>-</td>
<td>(7) &amp; (8)</td>
</tr>
<tr>
<td>Downconversion</td>
<td>(7) ⇒ (2)</td>
<td>LNA</td>
<td>(2) – (6)</td>
</tr>
<tr>
<td>Mixer</td>
<td>(8) ⇒ (2)</td>
<td>LNA</td>
<td></td>
</tr>
<tr>
<td>LO</td>
<td>No loopback</td>
<td>-</td>
<td>(5) &amp; (6)</td>
</tr>
<tr>
<td>Upconversion</td>
<td>No loopback</td>
<td>-</td>
<td>(3’), (4’), (5’), (6’), (7)</td>
</tr>
<tr>
<td>Mixer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test Attenuator</td>
<td>(7) ⇒ (1)</td>
<td>-</td>
<td>(1), (2), (7),</td>
</tr>
<tr>
<td>and Switches</td>
<td>(8) ⇒ (2)</td>
<td></td>
<td>(8)</td>
</tr>
</tbody>
</table>

Table 4.1 Test setups for the various RF blocks
Not only can individual circuits be tested but also a cascade test can be set up with the appropriate connections and nodes. Next we describe the various tests that can be run with the above setups such as gain, linearity, intermodulation distortion, and quadrature mismatch.

4.1.1 Gain

To extract the gain of an RF block, its input and output nodes have to be monitored. A one-tone test signal can be generated by the test-generation circuitry in the SoC and routed to any block along the loopback chain. Since the signals are measured at the boundary of the RF circuit, the block’s parameters are properly extracted thus eliminating any forms of performance masking that might appear in cascaded chains.

In the previous chapter, the detector’s accuracy (after some calibration) is shown to be within 5% to 10%, which translates to a gain prediction error of less than 1dB using only on-chip components. Fig. 4.2 depicts the gain measurement under a one-tone test. The predicted gain is then

\[
G_{\text{lin}} = f^{-1}(d_{\text{out}})/f^{-1}(d_{\text{in}}) \rightarrow A_2/A_1
\]

\[
G_{\text{dB}} = 20 \log (G_{\text{lin}})
\]

(4.1)

where \(f^{-1}(dc)\) is the mapping function relating the detector’s dc output to RF amplitude, \(d_{\text{in}}\) the output of the input detector, \(d_{\text{out}}\) the output of the output detector, \(A_1\) the amplitude of the input one-tone, and \(A_2\) the amplitude of the output tone.
Gain measurements for the mixer are a bit different as its inputs and outputs are at different frequencies. The detector can be leveraged to measure the amplitudes at the high frequency ports only, while a baseband sensor should measure the low frequency ports. Conversion gain can then be deduced from both sensors’ readings.

### 4.1.2 Compression Point

The compression point ($P_{1dB}$) of a CUT can be extracted following a one-tone test sweep. Using the same setup as the gain measurements described previously, the input signal amplitude is changed in discrete steps and the gain recorded. The input amplitude that results in a reduction of the gain by 1dB is then deemed the compression point. The discrete amplitude levels can be either supplied by the upconversion mixer for the transmitter tests, in addition to the loopback elements for the receiver tests. The gain
points gathered are used to construct the gain curve, as a piecewise fitting curve, and find
the input amplitude (and power) that results in a 1dB reduction in gain. Therefore,

\[
G_{\text{in}}[A_{1,x}] = f^{-1}(dc_{\text{out},x}) / f^{-1}(dc_{\text{in},x})
\]

\[
P_{1db} = (A_{1,y})_{\text{dBm}} \left| G_{\text{in}}[A_{1,0}] - G_{\text{in}}[A_{1,y}] \right|^{-1}
\]

where \(x\) is the iteration step and \(A_{1}\) the amplitude-swept input tone; \(A_{1,y}\) represents
the amplitude that results in compressing the gain, with \(y\) not necessarily being one of the
discrete set of points \(\{x\}\), i.e. it can be extrapolated between two discrete points.

### 4.1.3 Intermodulation Distortion

To test for intermodulation distortion, or more specifically IIP3, a two-tone test is
needed. Upon its generation by the transmitter baseband and upconversion, the two-tone
signal can be applied at various nodes in the system. Loopback elements have very good
linearity metrics of their own, by design, to keep the test signal from getting affected.

In Chapter 3, the response of the detector under a two-tone test is shown. While this
is not a pure dc signal, the basic premise here is that averaging that low frequency signal
is a task easily accomplished in the DSP following the ADC. The average corresponds to
a composite amplitude of the tones present in the signal. The two-tone signal without any
distortion should result a detector dc output mapped to exactly 1.414 times a single tone’s
amplitude (\(A_{1}\)). If a clean two-tone is assumed at the input, and the CUT gain (\(G\)) has
been tested for and known, then the expected output amplitude should map to exactly a
linear scaling of the input tones (\(\sqrt{2}A_{1} = \sqrt{2}GA_{1}\)). The emergence of third order
intermodulation distortion results in a detected amplitude in excess of the expected scaling. The difference between the detected and expected output amplitude is traced to the contribution of the new tones in the spectrum, the IM3 amplitudes \((B)\), as depicted in Fig. 4.3. That is, given the gain of the CUT \((G)\), the following holds

\[
\begin{align*}
    f^{-1}(dc_{in}) &= \sqrt{2}A_1^2 \\
    f^{-1}(dc_{out}) &\approx \sqrt{2}A_2(A_2 + B) \\
    B &= \frac{\sqrt{2}}{2} G \times f^{-1}(dc_{in}) \times \left( \frac{f^{-1}(dc_{out})}{G \times f^{-1}(dc_{in})} \right)^2 - 1
\end{align*}
\] (4.3)

where the averages of \(dc_{in}\) and \(dc_{out}\) are mapped through \(f^{dc}(dc)\) and \(A_1\) and \(A_2\) are the two-tone signals at the input and output respectively, and \(B\) the intermodulation distortion amplitude. The two-tone test signal power at the input needs to be below the compression point of the CUT for the above to provide accurate parameter extractions.

---

**Fig. 4.3** Two-tone test setup and IM3 measurement

\[
\begin{align*}
    \sqrt{2}A_1 &\rightarrow f\{dc_{in}\} \\
    \sqrt{2}A_2(B+A_2) &\rightarrow f\{dc_{out}\} \\
    \sqrt{2GA_1(B+GA_1)} &\rightarrow V_{inRFamp}
\end{align*}
\]
Again, the mixer is a special case. For example, a downconversion mixer’s IIP3 and IIP2 can be measured with the RF detector and a baseband detector. Two-tone test signals are also required from the test generation circuitry. The two-tone signals need to be carefully crafted to place either a second or third order intermodulation distortion product inside the output low-pass filter’s bandwidth. That component can then be detected by the baseband and used to quantify the IIP3 or IIP2 of the mixer in conjunction with the RF input amplitude, sensed by the RF detector, as in [50].

4.1.4 Quadrature Mismatch

Amplitude and phase mismatches in the quadrature modulation and demodulation results in incorrect reception and transmission of symbols. A primary cause of these mismatches is the local oscillator (LO). The high-frequency LO quadrature outputs, I and Q, are ideally of equal amplitudes and exactly 90° apart. The RF detector can be used here to obtain these mismatches by monitoring these two outputs.

Detecting amplitude mismatch is trivial as it can be directly spotted if there is a difference in the outputs of the detectors monitoring the I and Q signals. It can then be easily quantified. The phase between the quadrature signals, and hence the phase mismatch, can be measured by coupling both signals to a detector. Here, a differential detector is needed as well as differential quadrature LO, which is the most common implementation with the use of fully balanced differential mixers. By connecting one branch of the I signal and another from the Q signal to the differential input stage of the detector, a hybrid equivalent signal is sensed. A polar diagram that best demonstrates this
detection method is shown in Fig. 4.4. When sensing each path by itself, the respective detectors’ dc outputs correspond to the amplitudes of vectors $X$ and $Y$. If the detectors’ outputs are different, then $X$ and $Y$ are amplitude mismatched. However, when sensing the hybrid signal, the detector’s dc output corresponds to the resultant vector $Z$. Knowing the amplitudes of $X$, $Y$, and $Z$, the cosine law enables the extraction of the phase between the $I$ and $Q$ paths. Then phase mismatch is the deviation of that angle from $90^\circ$ as described in

$$
\begin{align*}
\text{amplitude mismatch} &= \frac{X}{Y} \\
\text{phase mismatch} &= 90^\circ - \frac{180^\circ}{\pi} \cos^{-1}\left(\frac{X^2 + Y^2 - Z^2}{2XY}\right)
\end{align*}$$

(4.4)

---

Fig. 4.4 Quadrature amplitude and phase mismatch measurement
4.1.5 Isolation and Feedthrough

Some elements in the transceiver should exhibit very good isolation between their ports. For example, a MOS switch in the off-state should ideally present infinite isolation. However, parasitics can couple part of the signal at one end of the switch to the other end. This is quite important in the loopback scheme discussed here, where the loopback element contains routing switches that on one end, at the transmitter interface, have high power signals and on the other end at the receiver interface, weak and low power signals. The RF detector can then be used to detect these RF components across an off-state switch. In such case, the one-tone test setup as depicted in Fig. 4.2 is used to quantify the extent of isolation, or lack thereof. A high-powered input test signal is preferred as it enables better visibility and detectability of an attenuated signal at the output.

Mixers also should exhibit a level of isolation to prevent the high-power LO signal from appearing at the IF or RF ports. This LO feedthrough is detrimental in both downconversion and upconversion mixers. In downconversion mixers, self-mixing due to the coupling of the LO signal to the input RF port results in an undesirable dc offset. To test for this, the setup shown in Fig. 4.5 can be implemented with the input RF port silenced (by disconnecting it from the other circuits, e.g. turning off the LNA and loopback) while the LO still drives its own port. If coupling is extremely low, then the RF amplitude detector will stay at its zero-RF dc value ($d_{hi}$). However, if the LO couples to the RF port, then it will excite the RF detector at that port that in turn will detect a high-frequency amplitude. Since no other RF input is present, then this detection quantifies the amount of LO coupling and subsequent self-mixing.
In the transmitter, the upconversion mixer’s LO feed-through (from the LO port to the output RF port) is problematic as it can fall in the vicinity or directly on top of the desired tone. One method to detect LO feed-through at the RF port of the upconversion mixer is to supply a low frequency tone at the transmit IQ baseband, and monitor the RF output with the RF amplitude detector. Without LO feedthrough, a single tone should be present at the output and the detector’s response will settle to a dc value corresponding to that signal’s amplitude. If an LO tone is also present, then the detector’s behavior resembles the two-tone case where its output is an oscillating signal at the frequency of the frequency offset between the LO and the upconverted test tone (see equation (3.6)
and Fig. 3.4). A visualization is shown in Fig. 4.6. For best LO visibility at the RF port, a smaller baseband test tone is preferred in order not to mask the coupled LO signal. Also it is assumed that no IQ mismatch is present in this setup as that will create also an image tone symmetric to the RF tone with respect to the LO frequency.

Fig. 4.6 Upconversion mixer test setup for LO feedthrough

4.2 Built-in-Self-Test Demonstration

In this section, we demonstrate the use of the detector in the previously described routines to characterize RF circuits and signal phenomena. Test-benches are first used to verify the usability of the detector in extracting some signal phenomena, irrespective of the circuitry it is attached to. This allows more controllability on the applied signals and provides a proof of the detector’s viability for such tests and signals. Then, actual circuits-under-test are monitored and their extracted parameters are compared to their
simulated ones. Two LNAs are built: one operating at 2.4GHz in 180nm CMOS and another operating at 60GHz in 90nm CMOS. The LNA is selected since it is one of the most critical circuits in the transceiver – and the RF detector is used to quantify its gain and linearity.

4.2.1 Detector Test-benches

Before connecting the detector to the RF circuits, we verify here the effectiveness of the detector in quantifying some of the signal phenomena that appear in transceivers, namely intermodulation distortion and phase mismatch. The signals are applied directly to the RF detector and the output is digitized using a 10-bit ADC and mapped to an amplitude measurement. In the following test-benches, the RF implementation of the amplitude detector (180nm) is used, as described in Chapter 3.

4.2.1.1 Intermodulation Distortion

To test the intermodulation distortion characterization method described in the earlier section, we implement a test-bench with the RF detector being supplied with a controlled two-tone signal. The amplitude of a single tone is set at 0.1V ($A_1$). Intermodulation components are added to the original signal and the output of the detector is observed and used to map to the corresponding amplitudes. As described earlier, the detector’s output in a two-tone test is a low frequency oscillating signal of which the average is used to compute signal amplitudes.
It is expected then that a clean two-tone stimulus to the detector should be interpreted as $1.4A_f$, or in this case 0.14V. Also, with increasing the injected IM3 component, the equivalent extracted amplitude (which is a function of the average dc output of the detector) should increase. The results are shown in Fig. 4.7, where the average dc output of the detector (at 1.08V) corresponds to around 0.14V (RF amplitude) when no distortion tones are present. Subsequent increases in the distortion tones (IM3) result in a decrease of the detector output and a resulting reciprocal behavior in the extracted amplitude.

Processing this information, we are able to extract the IM3 contribution given the offset created by the addition of that tone, in accordance with the previously established mappings in equation (3). Fig. 4.8 plots the predicted IM3 amplitudes versus the actual amplitudes (diagonal) showing very good correspondence.
Fig. 4.7 RF detector output and mapped output amplitude in response to a two-tone input with varying IM3 component

Fig. 4.8 Predicted versus actual IM3 amplitude
4.2.1.2 Quadrature Phase Mismatch

In this test, we demonstrate the extraction of the phase mismatch between two differential quadrature signals – I and Q. Three differential RF amplitude detectors are used where the differential I signal, Q signal, and the composite single-ended I/single-ended Q signal are fed to the first, second, and third RF amplitude detectors, respectively. These detectors will measure the amplitude components related to the I, Q, and the resultant vectors. In this test, no amplitude mismatch is considered and only changes in phase between the quadrature signals. Several detection runs are performed with phase mismatches between -10° and 10°. Referring to Fig. 4.4, the RF detector connected to the differential I signals records an amplitude $X$, while the one connected to the differential Q signals records a $Y$ amplitude. The composite amplitude $Z$ is also predicted. The phase imbalance is calculated using these amplitude measurements as described in equation (4).

Fig. 4.9 shows the results of the test runs with the predicted phases aligning with the actual phase shifts. The prediction error is within 1° for a 10-bit quantized detector output.
4.2.2 LNA as Circuit-Under-Test

In this section, we use the detector to characterize two LNAs at two different frequency bands. The one-tone and two-tone tests described in the earlier section are used to provide gain and linearity measurements. Also it is shown that the detector does not load the circuit-under-test. Detectors are placed at the input and output of the LNAs and their modes are selected depending on their dc output limits.

4.2.2.1 2.4GHz LNA

A 2.4GHz LNA is built in 180nm CMOS from TSMC and used with the RF amplitude detector. We will leave the implementation and circuit of the LNA for the next
chapter. The input one-tone is varied and the inputs and outputs of the LNA are observed and detected. The dc outputs of the detectors are quantized using an 8-bit ADC and are used to obtain the gain. Fig. 4.10 shows the results of the input sweep with the extracted gain matching to within 0.5dB of the real gain.

![Gain vs. Input Amplitude Graph](image)

**Fig. 4.10** 2.4GHz LNA gain extraction: actual versus predicted gain curve

### 4.2.2.2 60GHz LNA

A mm-wave LNA is implemented in 90nm CMOS from IBM [51]. The single-ended three-stage LNA implementation is shown in Fig. 4.11. Two mm-wave amplitude detectors with 8-modes (see Chapter 3 for description) attach to the input and output of the LNA. From Chapter 3, it was shown that this specific implementation of the detector presents around 1kΩ impedance and will therefore impact the LNA minimally. Fig. 4.12
shows the LNA s-parameters with and without the connected mm-wave detectors. It can be seen that the detectors do not cause degradation in the LNA’s original performance.

Fig. 4.11 60GHz LNA used for mm-wave BiST

Fig. 4.12 Effect of the mm-wave detectors on the LNA characteristics: with and without

One-tone and two-tone test sweeps are performed on the LNA with the detectors. The outputs of the detectors are quantized by an 8-bit ADC and the digital word is
mapped to an extracted amplitude measurement. After a number of one-tone iterations, the gain curve can be constructed and the $P_{1dB}$ point deduced. This is shown in Fig. 4.13 with the gain compression point highlighted.

Also two-tone test sweep is performed with a clean two-tone signal at the input of the LNA. The input and output detectors’ average dc output is again mapped to a resulting amplitude that can be used to extract the properties of the signals. Since the gain is known from the previous test, the intermodulation distortion BiST routine described earlier in the chapter is used to detect the emergence of third order intermodulations at the output of the LNA. Fig. 4.14 shows the results with the extrapolated curves for the simulated and predicted values. Table 4.2 summarizes the actual and extracted parameters of the LNA with the on-chip BiST matching the gain to within 0.3dB error, the 1dB compression point and IIP3 to within 0.4dB.

<table>
<thead>
<tr>
<th>LNA Parameter</th>
<th>Actual</th>
<th>Predicted</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gain, dB</td>
<td>10.14</td>
<td>10.45</td>
<td>0.3</td>
</tr>
<tr>
<td>$P_{1dB}$, dB</td>
<td>-9.73</td>
<td>-9.32</td>
<td>0.4</td>
</tr>
<tr>
<td>IIP3, dB</td>
<td>3.9</td>
<td>4.3</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Table 4.2  Actual versus predicted parameters for the 60GHz LNA
Fig. 4.13  Actual and predicted gain curve after one-tone test sweep

Fig. 4.14  Third order intermodulation extraction and IIP3 measurement
4.3 Summary

In this chapter, we expanded on the topic of Built-in-Self-Test as it applies to RF SoCs and described specification-based tests for the extraction of RF block parameters. We also placed the RF amplitude detector into the testing routine and used its capabilities to measure a number of circuit parameters such as gain, compression, and non-linearity, as well as amplitude and phase mismatch in quadrature signals.

These tests enable system self-awareness taking the system one step forward towards self-healing and self-calibration. Next chapter will describe the use of the amplitude detector in calibration routines and showcase a few circuit examples.
Chapter 5 | RF Built-in-Self-Calibration

The notion of cognitive radios came about a bit more than a decade ago, where it was suggested the radios should become powerful enough and smart enough to sense the outside world and environment and adjust their inputs and outputs accordingly. However, this deals only with the outside world, introspect cognition is also needed – an awareness of the radio’s own state – to allow it to adjust, or calibrate, its own internal components, especially now that the radio building blocks are becoming more volatile and prone to failure.

Calibration is usually done in the factory before an item is shipped and stored in memory, but this comes at an increased cost to the manufacturing and testing budgets. Self-calibration can be built into the system to be run at system start-up or at predetermined and periodic occasions, for example when a self-test deems that performance has noticeably degraded. The advantage of having a baseband DSP is a major enabler in making calibration self-contained and embedded within the system requiring no off-chip or off-system assistance.

For calibration to apply to RF blocks, they have to be designed and built with tunable elements in order to be programmed and tuned, on-the-fly. However, a key component here is self-awareness, on the block and system levels. A system can gain self-awareness
if proper testing and monitoring procedures are also part of its capabilities. In the previous chapters, we showed how a RF SoC can be made self-test-ready with the addition of loopback elements and most importantly accurate RF sensors. In this chapter, we take these capabilities and enhance them with self-adjusting capabilities of RF blocks, in line with the digitally-assisted RF design principles.

First, we describe how and why digital calibration is the most viable solution to analog impairments and highlight some of the self-calibration design requirements. Also, we show examples of digitally-assisted RF circuits and their calibration routines.

5.1 RF Block and System Self-Awareness

RF designers deal with an interesting challenge when building circuits in nanoscale CMOS: the integrated devices they use are now capable of offering good high-frequency performance only to suffer from very low reliability. The margin in which an RF circuit achieves its best performance is not only very narrow but also quite variable. The operating points that yield the optimal performance are not fixed; they are a function of many variables, including process, voltage and temperature. The increased susceptibility of individual transistors, the basic circuit building blocks, to sway away from their set characteristics makes the successful implementation of even the smallest and simplest of RF circuits an exercise in probability. While overdesign reigned in variability in older technologies, it simply cannot overcome the new challenges. One method is designing circuits with one degree of freedom, where the operating point of a block is not firmly set during design but fixed later at the factory production testing stage, usually through
component trimming and setting variables in some nonvolatile memory. However, this single degree of freedom, which is limited to a one-time fix, is still not inclusive of all operating conditions. To incorporate multiple fixes, factory calibration is no longer a valid option, and the burden is passed on to the only logical setting: the chip itself. The ultimate solution is assisted-operation whereby a circuit performs its functionality while being monitored and continuously steered to its optimal performance by supplying it the appropriate fixes. Assisted-operation is then heavily reliant on on-chip testability to instill awareness of the current circuit performance and operating conditions.

Self-awareness is therefore an inherited product of self-test and a necessary prerequisite for self-calibration. As was mentioned in the earlier chapters, BiST capabilities and techniques are at the core of any assisted circuit designs, especially for RF and analog. In the digital domain, BiST and self-calibration have been successfully applied to provide on-line thermal and power management adjustments. The push is to bring that on-line testability and flexibility to the RF and mm-wave domains. Hence, Design-for-Testability and Tunability (DfTT) should become the design method of choice for building firstpass RF SoCs.

RF circuit designs have traditionally opted for analog forms of compensation and feedback to reduce variability, with a multitude of techniques to remove dependency on voltage supply and temperature. However, these methods (mostly biasing circuits) are also implemented as analog circuits and suffer from the same adverse effects as the circuits they are supposedly adjusting. Another way to think of tunability is simply programmability – and what better method to program than in digital. Digital tuning of
analog blocks is an interesting alternative that offers a more robust alternative that can replace analog compensation and at the same time ease design requirements.

Analog compensation is a more self-contained method (i.e. a block and its compensation circuitry can ideally operate standalone). Digital compensation, on the other hand, requires the presence of a DSP to complete the compensation loop. That requirement is directly met in a radio SoC – and therefore, digital Built-in-Self-Calibration lends itself as a worthy topic of study and development.

5.1.1 Digital Solutions to Analog Impairments

The flexible processing capabilities of the SoC’s digital core can be put to use beyond the regular system operation. In the earlier chapters, the digital core is used to control the testing routines and compute different performance metrics. It can be further used to shield the system from failure caused by process and environment conditions. The advantages of the digital approach to monitoring and calibrating RF and analog circuits are many. Adaptive DSP algorithms can be programmed for block- and system-level calibration to achieve performance optimization. DSP solutions take only a fraction of the area of the blocks they are monitoring and would only need to run for a fraction of the time. Calibration routines can run at startup or at predetermined or idle times.

Several existing transceiver blocks can be re-used to make calibration possible, including the ADCs that form the interface to the digital core. The hardware overhead is then limited to the additional sensing and test circuitry. Also, the additional power
consumption can be limited only to the time calibration is running and the additional circuits turned off otherwise.

Portability and updatability of the digital approach are also major advantages. Digital circuitry almost always benefits from migration to newer CMOS processes, especially with the relative ease of digital synthesis. Moreover, the calibration algorithms, if implemented in software, can be easily updated allowing an increased level of flexibility, even providing companies with the ability to roll out firmware updates.

5.1.2 Enabling Built-in-Self-Calibration

The mixed-mode nature of RF blocks’ digital calibration borrows from the capabilities of built-in-self-test. For calibration to be run, test signals need to be applied to the circuit to be adjusted. The iterative process resembles the setups that can be used for one-tone or two-tone sweeps described in the earlier chapters on BiST. However, instead of dealing with the circuit as a black box and only analyzing its input and output signals, the sweep can also include changes to the circuit conditions. Essentially, this time testing is not used to extract the current performance as its end goal, but only a means to ensuring that the entire set of possible performance parameters is acquired and the best applied.

The self-calibration loop merges BiST and digitally-assisted design principles to force RF blocks into their optimal operating points. It is essentially an iterative process encompassing test, result comparison, and circuit correction. Therefore, looking at the basic loop in Fig. 5.1, the following actions are required at design time:
1. **Enable sensing:** Since the RF blocks’ signals are mostly at high frequency, they cannot be readily connected to the digital core for extraction of their properties and hence some type of feature translator is required. The translator takes in an RF signal and outputs a digitally friendly reading. The latter is the easily digitized and ready for DSP processing. One class of feature translators is discussed in this Manuscript (Chapter 3), the RF amplitude detector. Signal amplitudes can show a direct or indirect correspondence to circuit parameters and as such can be deconstructed and interpreted to provide a reading on a desired metric or a comparison between iterative calibration searches. Other sensing schemes can work at low frequencies to detect nonidealities in the system that are known to definitely present themselves at RF. An example would be charge pump current
mismatch in PLLs. It may be difficult to detect the resulting increase in phase noise due to that mismatch (emergence of reference spurs and possible increase in in-band noise). But it is definitely easier to observe and sense that low frequency mismatch at its source. Therefore, sensing needs to be enabled in circuits and functional blocks as part of a self-healing system.

2. **Design digitally-programmable circuits:** The transceiver blocks need to be designed with programmable or adjustable elements. The first step is to identify the weaknesses of the RF circuit under PVT variations. For example, threshold voltage mismatch between differential pairs can be simulated and its contribution to performance degradation quantified. More importantly is finding the most effective and suitable correction insertion points. Calibration structures can be designed to take on different values corresponding to the correction range and act as tuning knobs. The design task is to ensure that the correction range always has an optimal point of operation under PVT variations, as seen in Fig. 5.2. There may be different access nodes for the correction signals depending on whether they affect the operating point of the circuit (usually dc) or are present in the high-frequency operation. Circuit biasing is one of the major fixes that can be performed as it sets the dc operating point around which most of the ac characteristics center. Therefore, programmable dc biasing in the form of DACs are very popular tuning elements, to supply both variable voltages and currents, enabling adjustment of the circuit transconductances. Also, ac calibration can be
implemented in the form of continuous (fine) or discrete (coarse) tuning knobs, such as varactors or cap-banks or switchable loads.

Fig. 5.2 Digitally-assisted RF and mm-wave circuits implemented with wide operation flexibility and ranges always containing the optimal operation point

3. Formulate calibration algorithms: The cognitive part of the loop resides mainly in the DSP and implemented as calibration algorithms. When the system detects a change in the performance due to PVT variations, calibration can be run to revert the block to an appropriate performance level. This is exemplified in Fig. 5.3 where a change in PVT results in a shift of the operating space and hence the current operating point of the block. Upon detection of this change, maybe during periodically scheduled BiST, the calibration routine to search for the new best digital program word is started. The complete calibration of the system might
require a set of sub-calibrations working on individual blocks and an overseeing algorithm that manages all the routines. On the block level, prioritizing the calibration steps is essential and should follow appropriate circuit debugging methods. For example, a calibration algorithm for an LNA with both load tank resonance tuning and input match tuning elements might prioritize one or the other element as its primary fix or first order of action, such an LNA is discussed in the next section. This is also mirrored on the macro level, where system prioritization of block calibrations is also very important to achieve the required performance. Much like link budgeting described in Chapter 2, improving a single block’s performance might have a negative effect on other circuits and a possible negative net effect on the entire system. For example, the gain of an LNA cannot be driven to maximum as it might saturate and compress the blocks further along the chain. Therefore, calibration routines for single blocks need to be devised over which an arbiter algorithm makes the decision on what and when to run each.

5.2 Circuit-level Tuning

In this section we describe calibration cases using a number of RF blocks. First, a low-noise amplifier with input match and load tank calibration is presented along with its calibration algorithm and the use of the RF detector to sense its performance. Then, a mixer with various tuning knobs for gain and linearity is discussed. Also, a mixed-mode calibration scheme to correct IQ imbalances using the RF detector as a sensor is shown.
5.2.1 LNA Calibration

The LNA needs to appropriately receive a wanted weak signal, isolate it from other interferers, and properly amplify it to the input of the mixers. Since it directly interfaces to the antenna, the LNA has to ensure optimal power transfer by maintaining a perfect match. Moreover, on the output side, the load of narrowband LNAs is a resonant tank offering both gain and filtering centered on the band of interest. However, minor shifts due to parasitics on the input or output of the LNA can shift either or both of them off-center and away from acceptable performance. One of the most popular implementations, the inductively degenerated common-source LNA has an impedance matching network built with purely reactive components but presents a small-signal input impedance comprised of both reactive and resistive elements [52],

![Diagram of PVT Operating space with calibration steps](image-url)
\[ Z_{in} = j\omega (L_g + L_s) + \frac{1}{j\omega C_{gs}} + g_m \frac{L_s}{C_{gs}} \]  \hspace{1cm} (5.1)

where \( L_g \) and \( L_s \) are the gate and source inductors, \( C_{gs} \) is the input device intrinsic gate-source capacitance, and \( g_m \) its transconductance. At the resonance frequency, \( \omega_0 \), given by
\[
\omega_0 = \frac{1}{\sqrt{(L_g + L_s)C_{gs}}} \]  \hspace{1cm} (5.2)

the input impedance, \( Z_{in} \), is purely resistive and equal to
\[
Z_{in} = g_m \frac{L_s}{C_{gs}} \]. \hspace{1cm} (5.3)

To ensure maximal signal transfer, it is important to maintain both the resonant frequency at the desired signal band and the input match quality close to the antenna impedance. Process and temperature variations that alter the passive components’ typical values will influence the frequency response of the LNA. The only method to change back these values is to alter the reactive elements. In [30], the input impedance is changed by trimming the gate inductor, \( L_g \), using switches to short its segments at different locations, effectively creating a variable inductor. This approach is limited in a number of ways, including the need to design and verify custom inductors and the need to mitigate the effects of the finite switch on-resistance on the matching network.

Manipulating capacitance is in fact an easier approach to alter reactive values. One widely used example is varactor tuning. In [54], the authors suggest adding two varactors, one in shunt with \( C_{gs} \) and the other with \( L_s \). In their study on a differential 2.4GHz LNA, they show that the resonance frequency and the input match quality can then be
independently adjusted by varying the added varactors’ equivalent capacitance through a digitally-programmable bias generator (DAC). Based on the same idea, the load tank can be calibrated to maintain peaking at the required frequency also through varactor tuning of its output inductor [54], and also in [55].

Furthermore, it is shown that by applying a known signal at the LNA input and observing the output, the amplitude of the latter can provide an indication and measure of the LNA’s performance while being calibrated. The best LNA configuration is then the one that achieves the largest amplitude at the output – which is a direct result of a centered peaking frequency response at proper input match and output load resonance. Care should be taken however not to compress the LNA or the circuits following it. Therefore, self-test routines for linearity and compression points, presented earlier in Chapter 4, can set the limits of the calibration.

A 2.4GHz single-ended LNA based on [54] is implemented and shown in Fig. 5.4. The LNA is also augmented with the ability to be turned on or off by switching the cascade device’s bias between rails. This allows the LNA to be bypassed and shut down as required by other receiver blocks’ self-test and self-calibration routines. The LNA’s own calibration routine then starts by applying a signal from the loopback configuration. The input and output of the LNA are monitored by the RF amplitude detector presented in Chapter 3. The load tank is calibrated first, going through a linear search for the highest amplitude, and hence the lowest dc output of the detector. The optimal code is retained and applied for next phase of calibration. The same happens at each optimization level. Therefore, the calibration routine passes through the following steps:
1. **Enable Loopback**: create one-tone signal in transmitter and loopback to LNA input; monitor both the input signal (to ensure that it is stable) and the output signal of the LNA (the actual observable).

2. **Sweep \( C_d \)**: go through the load tank calibration codes (\( C_{d,\text{min}} - C_{d,\text{max}} \)) and save the optimal code (\( C_{d,\text{opt}} \)) corresponding to the lowest detector dc value (\( dc_{\text{out,min}} \)). At the end of the search, apply the optimal code and go to the next step.

3. **Sweep \( C_g \)**: sweep the calibration codes for shifting the resonance frequency of the input matching network; search for, and apply, \( C_{g,\text{opt}} \).

4. **Sweep \( C_s \)**: sweep the codes for matching quality adjustments; apply \( C_{s,\text{opt}} \) and end calibration.

---

Fig. 5.4 LNA with digital calibration for input match and output load tuning

---

---

124
5. **Disable loopback**: continue normal operation with new optimal LNA operating point \((C_{d,\text{opt}}, C_{g,\text{opt}}, C_{s,\text{opt}})\).

An example transient simulation displaying the progression of the calibration algorithm is shown in Fig. 5.5. The RF detector automatically changes modes when it reaches its upper and lower voltage limits; the higher the mode and the lower the dc output of the detector, the larger the amplitude of the signal is. Therefore, the final LNA setting will comprise of the set of varactor words that achieves the lowest possible detector dc output.

The gain of the final calibrated point can be directly extracted from the input and output RF amplitude detectors. These two detectors can keep on monitoring the regular operation of the LNA and if the system detects that the gain has significantly dropped from the last known good value, then calibration can be scheduled to run. Other modifications to the LNA circuit are possible to offer gain adjustment [56] and intermodulation distortion reduction [57].


5.2.2 **Mixer Calibration**

The mixer’s gain and linearity are of primary concern in a transceiver chain. The conversion gain in a mixer is defined for signals at different frequencies, at RF and IF, and can be measured using RF and baseband sensors. The intermodulation distortion products that are critical in mixers are not only of the third order (IM3) but also the second (IM2).

Self-test for these metrics is slightly different than that of other RF blocks as the mixer operates at two frequency domains. The RF amplitude detector presented earlier can be used on the RF and LO ports but not on the IF port – at least not for gain and linearity. It can still be placed on the IF port to measure the LO feethrough.

An active mixer mainly comprises of three sections: a transconductance part, a switching part, and output loads. The main signal enters the mixer through the
transconductance stage, whereas the local oscillator signal acts on the switching stage commutating the signal between the output loads. The various mixer parameters are then determined by one or more of these stages. For example, the mixer’s gain is simply a function of its input transconductance and output load. The IIP3 is also a function of the input transconductance. Therefore, adjusting the transconductance \( g_m \) of the input pair is a key component of manipulating both metrics. On the other end, the main source of IIP2 lies in the switching stage, mostly as threshold mismatch in the different commutating pairs. This mismatch creates an imbalance at the output node. One method to counteract this is to introduce an opposite set of mismatches to counterbalance the output. In [58], the counter mismatch is introduced as independently switchable output loads at either of the mixer outputs. The net effect is the cancellation of the original threshold mismatch. This technique unfortunately results in dc offsets, a main impairment in direct conversion receivers. The method suggested by [59] and [60] implements the counterbalance as independently biased switching pairs, thereby moving one step closer to the actual impedance \( V_{th} \) mismatch and its effect on the overdrive voltage of the transistors. By maintaining an equal overdrive voltage on the switching transistors, the net result resembles threshold-matched identically-biased switches, leading to the severe reduction in IIP2.

The double-balanced CMOS downconversion mixer implementation from [59] is shown in Fig. 5.6. The circuit’s gain adjustment includes transconductance and load tuning knobs. The first is achieved by pumping a variable current into the mixer through either the tail or pMOS current sources, controlled by \( V_{tail} \) and \( V_p \), respectively. The
second is simply a switched-resistor bank at both output nodes. The IIP2 adjustment knobs are the digitally tunable biases \( (V_s) \) at the gates of the switching pairs. A similar idea for IIP2 enhancement for passive mixers is presented in [61].

![Diagram of CMOS mixer with digital tuning knobs](image)

**Fig. 5.6** Double-balanced CMOS mixer with digital tuning knobs for gain and linearity calibration [59]

For this circuit’s calibration, the special case self-test mechanisms for gain and intermodulation distortion of mixers have to be engaged. To calibrate for gain, IIP3 and IIP2, the RF amplitude detector at the RF port is used to obtain the input amplitude whereas a baseband detector measures the resulting IF tones that fall within the low-pass filter bandwidth. When calibrating for gain, an RF one-tone signal is fed back to the mixer from the transmitter via the loopback element. The baseband detector observes the IF signal while the digital gain tuning codes \( (B_{\text{tail}}, B_p, \text{and } B) \) in Fig. 5.6 are changed. The search routine for the best set of codes can be done linearly or through a smart calibration engine employing least-mean square (LMS) adaptive algorithms in the DSP block.
IIP3 is also calibrated similarly however under a two-tone test while IIP2 is tuned using its own set of digital knobs ($B_{x,y}$ in in Fig. 5.6).

5.2.3 IQ Imbalance Calibration

Quadrature error is one of the major impairments affecting the performance of RF systems employing complex modulation schemes. The usual culprit is the local oscillator which fails to supply proper amplitude and quadrature matched signals to drive the mixers in either of the upconversion or downconversion paths. To achieve good IQ balance, designers have opted for layout techniques and new architectures that balance the outputs of the oscillator. However, variations are still largely probable and more so in nanoscale CMOS. An interesting calibration technique relies solely on the digital domain with no direct fix of the actual impairments. The technique, also known as “dirty-RF” [64], posits that knowledge of a transceiver’s analog impairments can be countered purely in the digital domain. That is, knowing the IQ imbalance of the transmit chain, a pre-distorted symbol is transmitted from baseband such that the introduced distortion and the inherent RF imbalance cancel out and a clean symbol is transmitted. The same also applies for receivers, where post-distortion of the received skewed symbols renders them ideal.

The first step is then to detect these imbalances. We have shown in Chapter 4 how using the detector to sense the differential quadrature outputs of the LO can enable amplitude mismatch detection and phase imbalance prediction accurate to less than a degree (see Fig. 4.9). A hybrid mixed-mode detection and calibration technique can then
be implemented, making use of the accuracy of the amplitude and phase mismatch detection with digitally-assisted analog compensation and DSP signal correction.

Using the amplitude and phase mismatch information obtained from self-test, the calibration algorithm in the DSP can correct for these mismatches in a number of possible ways. Several digitally-controlled analog correction knobs can be used to correct for the amplitude mismatch: LO phase-shift element tuning, mixer gain tuning (e.g. the mixer presented in the previous section), or alternatively tuning in baseband at the VGAs after (or before) the mixer. Once the gains of the $I$ and $Q$ paths are equalized, the phase mismatch remains.

Considering a demodulator input cosine signal at RF ($f_{RF}$) and phase mismatched ($\phi$) LO signals, the sampled $I$ and $Q$ signals after the low-pass filter are

$$
I[n] = \frac{1}{2} \cos(2\pi f_{IF} n) + d_i \\
Q[n] = \frac{1}{2} \sin(2\pi f_{IF} n - \phi) + d_q = \frac{1}{2} \sin(2\pi f_{IF} (n - r)) + d_q
$$

(5.4)

where $r=(\phi/2\pi f_{IF})$ is a fractional time delay, and $d_i$ and $d_q$ are dc mismatch terms [65]. The latter dc terms can be readily removed by the DSP. However, to complete the compensation, a fractional delay is needed on one of the channels. Fractional delay digital filters can hence be used to implement the phase calibration engine. One popular implementation of fractional-delay filters is the Farrow filter [66]. The Farrow filter is a time-varying FIR filter with fixed filter coefficients and a programmable delay component. Its structure is shown in Fig. 5.7 with $r$ as the fractional delay. It is essentially an interpolator that generates samples at fractions of the sampling period and
used in a number of diverse signal processing applications. It possesses many advantages including its guaranteed stability (as it is a feedforward architecture) and also low hardware complexity. Techniques to further reduce its hardware complexity have been devised to implement its coefficients as sum-of-products-of-two (SOPOT) therefore greatly easing the multiplication burden [67][68].

\[
x(n) \rightarrow C_1(z) \rightarrow C_2(z) \rightarrow \cdots \rightarrow y(n) = x(n-r)
\]

Fig. 5.7 Basic Farrow filter structure

Putting it all together, Fig. 5.8 shows the mixed-mode calibration setup for the receiver. A similar setup is possible for the transmitter pre-distortion. The IQ imbalances are extracted from the RF amplitude detector signatures at the LO. The amplitude mismatch is corrected by adapting the VGA of one of the paths whereas the phase mismatch is compensated in the digital fractional-delay filters. Fig. 5.9 shows an example of the effect of phase mismatch on the emergence of an image tone and its subsequent suppression by the Farrow filter in the digital baseband.
Fig. 5.8 Mixed-mode IQ imbalance compensation

Fig. 5.9 Test tone demodulation: quadrature phase mismatch before and after compensation

5.3 Summary

This chapter explored the built-in-self-calibration of RF transceivers and more importantly its requirements. In an RF SoC, the digital domain can be leveraged to compensate for analog and RF nonidealities by either correcting them on the circuit level or on the algorithmic level. The strength of the digital parts of the chip and the
capabilities that can be embedded in them allows for robustness enhancement measures to be put in place. This robustness is a product of self-aware blocks and systems that are capable of monitoring their own performance and mending its weaknesses. Self-test is therefore a vital part of calibration as it provides the metrics on which a calibration routine operates. The digital self-calibration loop is therefore enabled with efficient sensing, tunable circuits, and DSP algorithms. Examples of tunable circuits are shown with their calibration schemes using the self-test methodologies presented in earlier chapters.
Chapter 6 | On PLL Phase Noise Reduction

Phase noise is one of the most important metrics in PLLs, as it degrades the quality of the local oscillator tone and distorts the mixing process. Noise sources from each individual block of the PLL manifest themselves at the output. The loop dynamics affect how these noise sources contribute to the overall noise. The PLL is usually designed with a certain bandwidth that represents a careful compromise between many factors such as speed and noise. This bandwidth creates a set of filtering, or transfer, characteristics at each point in the loop, by which the noise sources get shaped. For example, noise from the reference crystal, the Phase Frequency Detector (PFD) and Charge Pump (CP) are low-pass filtered and are therefore the contributing elements of close-in phase noise. On the other hand, the VCO noise is high-pass filtered and dominates the far-off phase noise.

On-chip methods that measure timing jitter to infer phase noise have been embedded monolithically. However, an interesting phase noise spectrum on-chip measurement setup is discussed in [69], making use of a low-noise voltage controlled delay line (VCDL) and frequency discriminator to measure the phase noise at baseband. This type of detector may be used in a self-calibration routine with a programmable PLL to achieve the best phase noise performance. Digitally-programmable blocks can then be embedded in the loop to create a self-calibrated PLL. Apart from the overall phase noise, techniques to
measure individual blocks’ non-idealities have also been suggested, for example, CP current mismatch measurement and digital calibration [70][71].

One of the most critical non-idealities in the PLL is the CP current mismatch that brings rise to a number of undesired effects. This mismatch typically gives rise to reference spurs as the PLL forces a periodic correction of the current difference to maintain lock (by letting the lesser current on for more time to inject the same amount of charge). In fractional-N PLLs, an additional unwanted effect manifests itself in modulation-dependent CP-induced noise. This noise contribution, unlike the reference spur, appears at close-in frequencies that are not rejected by the low-pass filter transfer function at that point.

Therefore in this chapter, we look at methods to reduce Charge Pump current mismatch in itself, and also the modulation induced contribution by having a more favorable Sigma-Delta modulation type in the presence of mismatch. In the first section we show a programmable multi-mode digital Sigma-Delta modulator with the capability of changing orders and transfer functions. Second, third, and fourth order can be selected depending on the desired noise shaping. Even within the different orders, a number of modes are selectable where an optimal mode exists depending on the extent of CP current mismatch. The programmable modulator is implemented and verified on FPGA. The second section presents a charge pump (CP) design that simultaneously reduces current mismatch and sensitivity to process, voltage, and temperature variations. The self-biased and self-regulated circuit uses a dual feedback mechanism with a replica charge pump to dynamically stabilize and equalize the currents over a wide output voltage range under
varying operating conditions. The circuit is designed in 90nm CMOS and simulated using Monte Carlo analysis while sweeping the operating temperature to provide statistically relevant operating conditions. The proposed design achieves near perfect current matching at output voltages between 0.2V and 1V for temperatures ranging from -30 to 90 degrees Celsius and a power supply between 1.08V and 1.2V.

6.1 Reconfigurable Digital Sigma-Delta Modulators for Frequency Synthesis

At its simplest, an integer-N PLL fails to achieve wide tunability and fine resolution as it is limited by integer multiples (N) of a fixed crystal frequency. Additionally, its phase noise performance is degraded due to large values of N. Fractional-N PLLs solve this limitation by allowing for fractional multiplication of the reference frequency therefore enabling the use of a high frequency reference signal and much lower division values while still maintaining the required resolution. A fractional modulus is obtained as a time-averaged integer sequence. The most popular method to generate the integer sequence is the use of Sigma-Delta (ΣΔ) modulation as it offers two very desirable properties: randomization and noise shaping; whereby the integer sequence is not repetitive (reduces spurs) and quantization noise is shaped to higher frequency offsets (reduces noise near the synthesized tone) [69]. Furthermore, the noise shaping is a function of the modulator order – so given the requirements on output phase noise a proper order can be selected in conjunction with other important loop parameters such as loop bandwidth and reference frequency.
A \( \Sigma \Delta \) modulator (\( \Sigma \Delta M \)) for use in frequency synthesizers is all-digital – i.e. it performs the algorithm on a fixed digital input (\( L \), of size \( k \) bits) representing a fraction and outputs a randomized and noise shaped integer sequence (\( \text{OUT} \)) with an average of \( L/2^k \) and inherent instantaneous quantization noise (\( E \)). The input is shaped by an all-pass or low-pass signal transfer function (\( \text{STF} \)) while the quantization noise (\( E \)) is high-pass filtered by the noise transfer function (\( \text{NTF} \)),

\[
\text{OUT}(z) = \text{STF} \times L(z) + \text{NTF} \times E(z).
\] (6.1)

The high frequency quantization noise, shaped by \( 20 \times \Sigma \Delta \text{M}_{\text{order}} \) db/decade, enters the loop as phase noise and is later filtered to the output of the PLL by an equivalent low-pass transfer function, \( G(f) \).

In [73], we propose a multiple-order multimode reconfigurable \( \Sigma \Delta \) architecture able to offer noise shaping flexibility to meet phase noise requirements of various wireless standards. The modes of operation are carefully selected to be 1) easily implemented in hardware and 2) offer increased immunity to loop non-idealities. Although, a few studies discuss the first issue [84][85], the novelty of this design mainly stems from its consideration of the second.

The contribution of the \( \Sigma \Delta \text{M} \) quantization noise to the output phase noise can be modeled as shown in Fig. 6.1. The frequency divider in the feedback loop of the PLL acts as a digital accumulator transferring the modulator quantization error to phase error (\( \Phi_{\text{SDMN}} \)) before it enters the loop – to get low-pass filtered by \( G(f) \) [73]. Modulation-dependant second order effects such as Charge Pump mismatches (static and dynamic),
Phase-Frequency Detector dead zones, and divider delay also affect the output phase noise [75][76][77]. However, the most dominant of these noise sources is the charge pump (CP) static mismatch (between the up and down current, usually given in percent mismatch $\rho_{\text{mismatch}}$)[76] – hence only this modulation-dependant noise source is considered here and shown in the bottom part of Fig. 6.1. This mismatch results in high frequency noise being folded down to lower frequencies thus adding to the close-in phase noise. The added mismatch noise, $\Phi_{\text{CP}SDM}$, is directly proportional to the distribution of the phase error, $\Phi_{\text{SDMN}}$, and consequently to that of the $\Sigma\Delta M$ output. Therefore, the output phase noise attributed to the $\Sigma\Delta M$ is the aggregate of both $\Phi_{\text{SDMN}}$ and $\Phi_{\text{CP}SDM}$ transferred to the output.

Fig. 6.1  Output phase noise due to Sigma-Delta modulation: contribution from the modulator and the charge pump static gain mismatch (example spectra shown for a third order SDM with a noise shaping of 60db/decade)
In our analysis, we try to find the most suitable modulator architectures across the most commonly used orders, i.e. second, third, and fourth. Higher order modulators offer more randomization, better shaping, and more output levels. The downside is that with the increase in output levels, their distribution widens and their amplification of CP mismatches increase. A single fixed ΣΔM architecture with a certain distribution of output levels fails to offer the best phase noise performance when random nonlinearity magnitudes in the analog blocks exist.

There are two major classes of high order ΣΔMs: single loop modulators and cascaded modulators. The latter, known as MASH, is a very efficient randomizer and noise shaper yet has a high number of output levels and a fixed response. Single loop modulators, on the other hand, allow for a large degree of flexibility with a number of possible implementations; however, special attention should be taken to ensure stability of the loop. In this study, we choose to design a single loop feedforward architecture of which Fig. 6.2 shows a 4th order implementation. Its flexibility stems from the parameters that can be changed such as the coefficients (Bᵢ), the choice of integrator (Hᵢ), and the number of bits in the quantizer.

Fig. 6.2 Feed-forward single-loop SD Modulator
We define the open loop gain as

$$F = B_1 H_1 + B_2 H_1 H_2 + B_3 H_1 H_2 H_3 + B_4 H_1 H_2 H_3 H_4$$  \hfill (6.2)$$

where $H(z)$ is either a delaying ($1/(z-1)$) or nondelaying integrator ($z/(z-1)$).

By simply setting $B_4$ to zero, a 3rd order modulator open loop gain can be achieved. Doing the same for $B_3$ makes a 2nd order open loop gain. The output can then be expressed as,

$$OUT(z) = \frac{F}{1+F} \times L(z) + \frac{1}{1+F} \times E(z).$$  \hfill (6.3)$$

An important issue is the stability of the modulator which can be assessed by the location of the poles of its transfer functions. Therefore, special care should be taken to insure that the poles are inside the unit circle (in $z$-domain); and this is a function of the feedforward coefficients as well as the integrators used. In this study we are interested in stable configurations with digitally implementable designs (i.e. power of 2 coefficients are highly desirable as they are implemented with simple bit shifting). Therefore, preliminary analysis on stable integrator configurations and their corresponding stable set of coefficients is performed. An example is shown in Fig. 6.3 for 3rd order single-loop ($B_4$ set to 0) where the range of stable coefficients, $B_1$ and $B_2$, with $B_3$ set to 0.5 when the integrator configuration is Delaying – Nondelaying – Nondelaying ($H_1 H_2 H_3 = DNN$)[80]. The enclosed region represents the coefficient pairs that yield stability and the preferable $B_1$ and $B_2$ pairs are highlighted by dots. Further assessment of the possible stable implementations enables the exclusion of those with undesirable peaking in their transfer functions due to their poles’ location (close to the unity circle).
6.1.1 Analysis

A number of stable and digitally implementable modulators is selected in each of the orders. MATLAB/Simulink $\Sigma\Delta M$ models are created and simulated in order to gather data for comparison and performance evaluation. The models are verified to work for different bus widths and the entire fractional range from 0 to 1 [83]. In order to reduce the limit cycle effect in the modulators, the input LSB is set to 1 to approximate an irrational fraction since rational fractions such as 0.25, 0.5 and 0.75 cause a very short repeating output sequence manifesting as spurs [82].

The noise model described in Fig. 6.1 is used to benchmark the various modulators with respect to their respective noise shaping abilities and the extent of their non-ideality (CP mismatch) contribution to output phase noise under different loop bandwidths and reference frequencies. Furthermore, we make use of the integrated phase noise as a figure...
of merit in comparing the ΣΔMs. Extensive simulations covering the entire fractional range of 0 to 1 are performed and then from the resulting output sequences, output phase noise can be computed in the frequency domain. The loop bandwidth is varied (hence $G(f)$) as well as the reference frequency ($F_{\text{ref}}$). As 3$^{\text{rd}}$ order is more common and 2$^{\text{nd}}$ order does not offer good noise shaping, we prioritize the analysis as such: 3$^{\text{rd}}$, 4$^{\text{th}}$, and then 2$^{\text{nd}}$ order. Additionally, we opt to arrive at modulator configurations that are architecturally similar to make use of efficient hardware sharing in their implementation.

6.1.1.1 Third order modulators

Among a number of possible implementations, three 3$^{\text{rd}}$ order ΣΔM modulators fare well with respect to CP mismatch. Fig. 6.4 shows the integrated phase noise for these modulators (labeled as (a), (b), and (c)) with a reference frequency of 40MHz and a loop bandwidth of 1MHz. It can be seen that each of the modulators is optimal at a certain range of CP mismatch percentage. It is also noted that the three modulators have a similar integrator configuration ($DNN$) but varying feedforward coefficients – see Table 6.1. More extensive simulations are performed while varying the reference frequency, loop bandwidth, and CP mismatch. This results in a more generalized view of which modulator is optimal given the ratio of the reference frequency to loop bandwidth ($F_{\text{ref}}/BW$) and the percentage mismatch, as shown in Fig. 6.5 and in [80]. Modulator (a) is identical, in terms of noise shaping and output levels distribution, to a cascaded MASH 1-1-1; and it can be seen how its integrated phase noise is degraded with slight CP mismatch, as also confirmed in [79].
6.1.1.2 Fourth order modulators

The same procedure is applied to fourth order modulators. However, this time we build on the established architecture similarities of the third order modulators (having $H_1H_2H_3$ as DNN) to find optimal 4th order architectures. We construct simulation models for two viable integrator configurations ($H_1H_2H_3=DNN$ and $H_4$ as delaying or non-delaying).

Again three configurations are optimal depending on the $F_{\text{ref}}/BW$ ratio and CP percentage mismatch. These modulators are labeled (d), (e), and (f) in Fig. 6.5(right) and Table 6.1 shows their configuration. Modulator (d) has an NTF and output levels distribution identical to that of MASH 1-1-1-1.
Fig. 6.5 Optimal 3rd (left) and 4th (right) order modulators as a function of $F_{ref}$, loop bandwidth, and charge pump mismatch percentage

6.1.1.3 Second order modulator

Similar simulation and analysis result in only a single second order modulator being optimal for all values of $F_{ref}/BW$ and mismatch. This modulator is labeled (g) in Table 6.1.

<table>
<thead>
<tr>
<th>Modulator</th>
<th>Order</th>
<th>$H_1H_2H_3H_4$</th>
<th>$B_1B_2B_3B_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)       * 3</td>
<td>DNN-</td>
<td>1-1-1-0</td>
<td></td>
</tr>
<tr>
<td>(b)       3</td>
<td>DNN-</td>
<td>1-0.5-1-0</td>
<td></td>
</tr>
<tr>
<td>(c)       3</td>
<td>DNN-</td>
<td>1-0.5-0.5-0</td>
<td></td>
</tr>
<tr>
<td>(d)       ** 4</td>
<td>DNNN</td>
<td>1-1-1-1</td>
<td></td>
</tr>
<tr>
<td>(e)       4</td>
<td>DNNN</td>
<td>1-1-1-0.5</td>
<td></td>
</tr>
<tr>
<td>(f)       4</td>
<td>DNNN</td>
<td>1-1-0.5-0.5</td>
<td></td>
</tr>
<tr>
<td>(g)       *** 2</td>
<td>DN--</td>
<td>1-1-0-0</td>
<td></td>
</tr>
</tbody>
</table>

*NTF similar to MASH 1-1-1 ** NTF similar to MASH 1-1-1-1
*** NTF similar to MASH 1-1

Table 6.1 Optimal feedforward single-loop modulators and their configurations
6.1.2 Proposed modulator

6.1.2.1 Architecture

Based on the results for the optimal modulators in the previous section, we propose a hardware implementation that covers all the modulators. The architecture makes use of hardware similarities found in the integrator configurations as well as the simple single-bit shift to implement a coefficient of 0.5. Fig. 6.6 shows this architecture with the various inputs, outputs, and control bits.

Integrators are made up of adders and memory elements where the output of the adder is connected back to one of its inputs through the memory element (thus introducing a delay). A delaying integrator uses the output of the memory element as an output whereas a non-delaying integrator uses that of the adder. From Table I, \( B_1 \) is always 1 whereas the remaining coefficients are either 1 or 0.5 (when used), or 0 (when off). In the proposed design, \( B_2 B_3 B_4 \) are used to control single-bit shifters when a coefficient of 0.5 is needed. Two control bits, namely \( o_3 \) and \( o_4 \), define the order of the modulator by connecting and disconnecting the outputs of the 3\(^{rd}\) and 4\(^{th}\) integrators. For 4\(^{th}\) order operation, both \( o_3 \) and \( o_4 \) should be asserted thus connecting the shifters to the four-operand adder; for 3\(^{rd}\) order operation, \( o_4 \) is set to 0 and disconnects the last integrator. Likewise, for 2\(^{nd}\) order operation, both \( o_3 \) and \( o_4 \) are set to 0. Multiplexers may be used to provide shifting and order reduction. These multiplexers can be controlled by the coefficients, \( B \), to pass either the integrator’s original result or its shifted (divided-by-2) version in addition to being controlled by either \( o_3 \) or \( o_4 \) to pass zeros when a lower order mode is used. The input
$(L)$ is connected to the subtractor that takes its other operand from the feedback output bits ($OUT$). For $n$-bit adders and flip-flops, the fractional input $L$ is $k=(n-4)$ bits wide.

### 6.1.2.2 Resource Usage Reduction

An $n$-bit realization of the architecture uses adders/subtractors, memory elements, and multiplexers that are $n$-bit wide. By employing inter-stage error masking [86], an $n$-bit realization can be obtained with only the first integrator, $H_1$, being $n$-bits wide while reducing the width of the subsequent integrators. This introduces further quantization errors; however, these errors can be masked below $E(z)$ given an appropriate minimum width for the subsequent stages, $n_x$, satisfying the following inequality

$$2^{l-n_x} \sum_{i=1}^{d} \left( \frac{2^{l-k-x+4}}{(2\pi)^2} \right)^{i-1} < 1 \quad \text{for } x \in \{2,3,4\}, n_1 = k + 4 \quad (6.4)$$

where $k$ is the fractional width and $n_1 = n = k + 4$ is the width of the first integrator.

For example, a 20-bit implementation ($k=16$) can be reduced from an all-20-bit to a 20-16-15-13 for $n_1-n_2-n_3-n_4$, $-10\%$ reduction in resource usage as shown in Table 6.2. It is also seen that the adaptable/reconfigurable architecture presents no substantial overhead in terms of resource usage ($-20\%$) and maximum operation speed as compared to a simple $4^{th}$ order single loop design all while offering 7 modes of operation.
Fig. 6.6 Proposed multimode reconfigurable ΣΔM architecture

<table>
<thead>
<tr>
<th>binary</th>
<th>integer</th>
<th>binary</th>
<th>integer</th>
<th>binary</th>
<th>integer</th>
<th>binary</th>
<th>integer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>8</td>
<td>0100</td>
<td>4</td>
<td>0000</td>
<td>0</td>
<td>1100</td>
<td>-4</td>
</tr>
<tr>
<td>0111</td>
<td>7</td>
<td>0011</td>
<td>3</td>
<td>1111</td>
<td>-1</td>
<td>1011</td>
<td>-5</td>
</tr>
<tr>
<td>0110</td>
<td>6</td>
<td>0010</td>
<td>2</td>
<td>1110</td>
<td>-2</td>
<td>1010</td>
<td>-6</td>
</tr>
<tr>
<td>0101</td>
<td>5</td>
<td>0001</td>
<td>1</td>
<td>1101</td>
<td>-3</td>
<td>1001</td>
<td>-7</td>
</tr>
</tbody>
</table>

Table 6.2 Resource usage comparison

<table>
<thead>
<tr>
<th></th>
<th>LUT</th>
<th>FF</th>
<th>Fmax</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>20-bit</strong> 4th order single-loop</td>
<td>102</td>
<td>65</td>
<td>130MHz</td>
</tr>
<tr>
<td><strong>20-20-20-20</strong> proposed multimode</td>
<td>136</td>
<td>64</td>
<td>122MHz</td>
</tr>
<tr>
<td><strong>20-16-15-13</strong> proposed multimode</td>
<td>124</td>
<td>61</td>
<td>123MHz</td>
</tr>
</tbody>
</table>

6.1.2.3 Implementation and Results

The proposed design is verified by running a 20-bit (20-16-15-13) VHDL implementation on a Stratix II EP2S60 DSP Development Board from Altera Corporation
in various modes of operation and collecting the digital output samples to be processed in MATLAB. The different orders of noise shaping are visible in the output PSD as well as the flexible noise transfer functions in each order. Fig. 6.7 shows the results of the runs of modulators (a) through (f) – modulator (g) is not shown due to space constraints. The theoretical noise transfer functions (dotted lines in Fig. 6.7) are also plotted and show direct correspondence with the actual results. The 40, 60, and 80dB/decade noise shaping are readily verified for the 2\textsuperscript{nd}, 3\textsuperscript{rd}, and 4\textsuperscript{th} order modulators respectively.

![Fig. 6.7 Output spectra of third and fourth order modulators (a)-(f). The theoretical noise transfer function for each modulator is plotted (in dotted line) on top of the PSD [vertical units: db/bin, horizontal: Hz]](image)

6.2 Dynamic Self-Regulated Charge Pump Circuit

A number of the non-idealities affecting PLL performance relate to the CP as it supplies the charges to the loop filter to regulate the voltage input of the VCO. Charge
sharing, charge injection, transient response, and current mismatches all contribute to noise at the output. The latter, unequal sourcing and sinking currents, exhibits itself in increased reference spurs and in the case of fractional-N PLL’s, an increased close-in noise floor at the PLL output due to Sigma-Delta noise folding phenomena. Therefore, it is highly desirable to achieve matching $I_{UP}$ and $I_{DN}$ currents. However, with reduced power supplies and feature sizes, circuits built in nanometer scale technologies suffer from an increased susceptibility to performance degradation with process, voltage, and temperature variability, collectively named as $PVT$.

Here we describe a CP with simple dynamic dual feedback compensation to provide very good current matching over a wide range of operating conditions.

### 6.2.1 Proposed method and circuit

Fig. 6.8 shows a conventional CP design and its current characteristic as a function of output voltage. Several methods have been proposed to reduce the effect of the output voltage ($V_{CPout}$) on the sourcing ($I_{UP}$) and sinking ($I_{DN}$) currents mainly due to the channel-length modulation phenomena of the non-identical PMOS and NMOS devices, especially in modern short channel technologies. The classical approach is to use long channel devices to increase the output impedance of the pump at the expense of speed and area, as well as increased parasitics. Alternatively, gain-boosting [87] and cascoding have been used to increase the pump’s output resistance. On the other hand, op-amps have been employed to provide negative feedback compensation [88][89]. However, such techniques come with their own set of issues such as op-amp stability, offset, power
consumption, noise and increased area. Moreover, the CP current characteristics are affected not only by the CP output voltage but also process mismatches and temperature fluctuations. The same also applies for the performance of any added compensation circuitry thus making current matching a difficult issue to tackle.

The proposed CP circuit uses dual feedback to reduce the effects of channel length modulation and to force current matching over PVT variations. Fig. 6.9 shows the circuit implementation where source switching is used (SW1 and SW2) and the sourcing current ($I_{UP}$) and sinking current ($I_{DN}$) flow through $P1$ and $N1$ transistors, respectively. Current mirrors composed of $N2-7$ and $P2-7$ bias the transistors at the pump output branch. The main biasing branch is comprised of $N10-11$ and $P10$. The first feedback is built into the

Fig. 6.8 Conventional charge pump and current characteristics
individual CP whereas the second is enabled through the incorporation of an identical replica CP.

![Proposed charge pump circuit with dual-feedback](image)

**Fig. 6.9** Proposed charge pump circuit with dual-feedback

First, to increase the matching between $I_{UP}$ and $I_{DN}$, the CP output is fed to two feedback transistors, $N4$ and $P4$ [90]. This feedback, shown in Fig. 6.10a, reduces $I_{DN}$ and $I_{UP}$ at low and high CP output voltages, respectively, making them match over a wide range of output voltages without the use of long channel devices that slow the transient response or feedback operational amplifiers that add noise to the circuit.
Second, since the threshold voltages are not decreasing at the same rate as the power supply, the CP output current can be held on for a smaller output voltage range before either of the branches goes into the triode region hence reducing the output current. To counter that, we propose a boosting technique for low and high output voltages. Again, the CP output is fed to two simple inverting circuits \((N11/P12 \text{ and } N13/P16)\) that control the biasing of \(V_{nBias}\) and \(V_{pBias}\) through \(P13\) and \(N14\), respectively. \(P13\) increases \(V_{nBias}\) at high CP output voltage to boost \(I_{UP}\) and, conversely, \(N14\) decreases \(V_{pBias}\) at low CP output voltage to achieve the same for \(I_{DN}\). The action of these two inverters on the biasing of the current mirrors widens the matching range as depicted in Fig. 6.10a and demonstrated in simulation in the next section.

Fig. 6.10 Feedback concept of the proposed charge pump
While this reduces the effect of channel-length modulation and results in a flatter characteristic, it does not account for the non-deterministic PVT variations that compromise the matching. When the currents are not matched, the CP current characteristic shifts and the currents are equal for a single output voltage, $V_{cross}$, as shown in Fig. 6.10b. To re-establish matching, the replica CP is used to detect this cross-over voltage. The replica circuit is always-on and its output voltage will settle to $V_{cross}$ to ensure equal currents in its output branches. As depicted in Fig. 6.10b, when $I_{DN}$ is larger than $I_{UP}$, the replica CP output voltage settles to a low crossover voltage whose voltage complement $V_{fb}$ – obtained by a simple inverting buffer circuit N8-9 and P8-9 [91] – drives the second set of feedback transistors, N7 and P7. The high voltage $V_{fb}$ alters the triode-region N7 and allows an increase in the current being mirrored to the sourcing branch, therefore increasing $I_{UP}$. Simultaneously, P7 is made less conductive resulting in a reduced $I_{DN}$. Similarly, the complimentary action occurs when $I_{DN}$ current is larger than the $I_{UP}$ current. The compensation results in equal $I_{UP}$ and $I_{DN}$ and dynamically tracks the fluctuations mainly due to voltage (power supply) and temperature variations.

### 6.2.2 Simulation results

To test and verify the proposed circuit and matching technique, we design and simulate the charge pump circuit in 90nm CMOS technology at 1.2V power supply. First the individual CP is demonstrated to achieve a wide matching region under typical PVT conditions (typical N/PMOS, 1.2V, and 30ºC). The mismatch is shown to be less than 0.1% over a 0.18 – 0.98V output range (without the range extension effects of P13/N14
the current would match over the 0.3 – 0.8 V output voltage range only. This is shown in Fig. 6.11 versus the conventional CP from Fig. 6.8.

![Proposed CP current matching across output voltage versus a typical CP](image.png)

Fig. 6.11  Proposed CP current matching across output voltage versus a typical CP

At non-typical process corners, the deviation between $I_{UP}$ and $I_{DN}$ currents is significant. Fig. 6.12 shows the simulation results under slow NMOS and slow PMOS process corner and 60°C where the dashed lines represent $I_{DN}$ and $I_{UP}$ currents without replica feedback along with their percentage mismatch (~12%). When the replica CP and inverting buffer are included, both $I_{UP}$ and $I_{DN}$, denoted by the solid lines in Fig. 6.12, match and their maximum mismatch in the 0.2-1V output voltage range is less than 1%. 

154
This also holds true for wide temperature (-30 to +90°C), process/mismatch (±3σ), and supply voltage (90% to 100% Vdd) variations. Monte Carlo simulations are used to obtain a statistical measure of the effectiveness of the proposed technique and CP design while sweeping the operating temperature and supply voltage. The temperature is allowed to vary from -30°C to 90°C in 30 degree increments while the supply voltage is scaled between 1.08V and 1.2V. Two-hundred simulations per temperature-voltage point are performed while allowing for process statistical variations pertaining to the 90nm CMOS technology. Fig. 6.13 shows a sample mismatch distribution of 200 simulations at 30°C and 1.2V supply voltage for the conventional CP, the proposed CP with only local feedback, and the full proposed implementation with replica global correction. It can be seen that the conventional implementation cannot match well over process changes and similarly does the partial implementation. However, with the global correction through the replica, matching is maintained throughout the process and device mismatch.
variations, to within 1%. Fig. 6.14 presents the results of all the simulations with and without feedback at various temperatures. It can be seen that the proposed implementation achieves a very good match over PVT averaging to less than 2%.

![Bar chart showing current mismatch for different temperature settings and correction feedback methods.](image)

Fig. 6.13 Monte Carlo simulation results for current mismatch at T=30
In this chapter, we looked at one of the most critical mismatches in a PLL that creates noise at both far-off as well as close-in offsets from the carrier. The charge pump current mismatch therefore needs to be reduced to eliminate reference spurs and for that we have proposed a self-calibrated charge pump architecture. Moreover, this mismatch is also agitated by the modulation noise in fractional-N PLLs created by the Sigma-Delta modulator. Therefore, we show that there exist several architectures that offer optimized modulation-induced noise depending on a number of loop parameters such as reference frequency, bandwidth, and CP mismatch. For this, we proposed a single implementation.
that is capable of flexible modulation order and characteristics that can be changed to reduce the CP mismatch contribution to output phase noise.
Chapter 7 | Conclusions

The technology breakthroughs resulting in the successive scaling of CMOS transistors has kept Moore’s law alive for more than half a century. An important effect of physical dimension reduction is the increase in attainable operating frequencies. Therefore, devices are not only smaller but also faster, fast enough to surpass the RF spectrum and even the millimeter-wave region. This allowed CMOS to contend for integrated circuits for wireless connectivity applications. With the proliferation and surge of demand for ubiquitous computing and connectivity, the customer pool has expanded tremendously making wireless (devices, applications, services, etc…) one of the fastest growing markets worldwide. The new challenge then becomes to take the cost-effective and mass-production-ready CMOS beyond memory and logic into the “More than Moore” regime. This represents a new dimension in integration, whereby hybrid mixed-mode systems co-exist on the same silicon substrate, essentially combining memory, logic, power, analog, and RF circuits into one: a radio System-on-Chip.

Integrating RF and mm-wave capabilities with powerful digital processing offers a great deal of possibilities, if only it were an easy task! The challenges turn out to be as great as these possibilities. The inherent increase in variability of nanometer CMOS device performance translates into extreme variations of circuit functional metrics.
Coupled with the tighter requirements for new applications and standards, the percentage of working parts that also pass the specifications greatly diminishes. The yield problem is more pronounced in RF SoCs as the radio frequency circuits that occupy only a part of the area are much more vulnerable to failure thereby setting the passable limits of the entire system. The seemingly unwieldy RF circuits do have an optimal operating region; the only downside is the narrowness of that region making minor shifts in operating conditions manifest as extreme roll-offs in performance. However, that operating region can be still recuperated – although demanding direct intervention. To regain lost performance, the first step is to know what the current state of operation is. Therefore, testing is the first step towards obtaining this knowledge. Techniques for testing RF circuits are then of primary importance. External testing is quite prohibitive, time- and cost-wise, especially in highly complex systems employing a mix of digital, analog, mixed-signal, and RF. Design-for-Test provides a partial solution to this difficulty by enabling test structure on-chip embedding, for increased visibility and observability. However, external testers are still required which limits testing and verification to the lab.

In this manuscript, we highlighted the requirements for migrating the test completely on-chip and allowing it to be portable. The portability property carries with it some interesting possibilities of testing a system on-the-fly and while in actual operation, where changes in operating conditions are the most probable and representative. This equates to building functional blocks in conjunction with testing blocks at the very onset of the design process. However, the desire is to include more functionality and not more testability, putting a reasonable limit on the overhead associated with on-chip test. To this
end, we propose a test sensor that lends itself to accurate parametric extraction of various RF and mm-wave blocks and, at the same time, to non-invasive and minimal-overhead integration. The sensor, an RF amplitude detector, translates a high frequency signal into a corresponding low frequency or dc reading. That reading can be easily assessed by the digital parts of the SoC and helps in quantifying a signal’s properties and a block’s performance. The proposed RF detector benefits from high conversion gain (sensitivity) and wide dynamic range (extent of amplitude detection) while being frequency non-specific. Three implementations are compared to available solutions in literature whereby our designs position themselves as front contenders for true on-chip BiST. An enhanced loopback BiST architecture is also proposed making use of a signal loopback element and multiple sensing nodes to perform a number of parametric tests. The complete system should be able to locally generate test signals, sense the response to these signals, and interpret the signatures of the sensing results. These steps are all possible within the boundaries of an SoC, and with very little overhead. The RF amplitude detector links the RF and digital at its ends.

Several block metrics can then be inferred directly or indirectly from simple amplitude measurements. The test setups are discussed for a number of blocks from LNAs, mixers, local oscillator, and quadrature modulators and demodulators. Methods to quantify gain, linearity, intermodulation distortion, isolation, and quadrature mismatch are explained and example cases demonstrated for both micro- and millimeter-wave spectrum.
The ultimate goal is not only test but actual performance enhancement. There is a certain limit to making an RF circuit rigid and highly stable. This limit has been pushed to extents demanding unreasonable design effort while still not guaranteeing proper functionality across all cases. Surprisingly, the answer is exactly the opposite: flexibility. Flexible RF circuits are not overdesigned to perform at their best operating point but actually loosely designed with enough wiggle room. The margin is relaxed in anticipation of assisted-operation, i.e. an active mechanism will regulate the operating point to match the operating conditions resulting in performance enhancement. The testing techniques we presented will enable the system to become self-aware and therefore know its location within the operating margin and work its way towards the optimal point inside that margin. With flexible digitally-assisted RF circuits, the optimization can run on the more robust and highly-efficient digital parts. A few examples for digital tunability of RF circuits are presented including LNA, mixer, and quadrature modulator. The highlighted techniques and test strategies can then be extended to a number of other circuits. The phase noise of PLLs is also addressed in the context of reducing the charge pump mismatch contribution to output noise, both in integer-N and fractional-N PLLs.

With the newer CMOS processes showing no sign of variability restraint, a successful self-calibration architecture is tantamount to a successful chip implementation. Therefore, the importance of self-test and self-calibration for RF SoCs cannot be underestimated and has to be employed as a new multifaceted design paradigm employing techniques and best practices from across the board.
References


165


Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan, www.tsmc.com


