Mohammed Ismail

Delia Rodríguez de Llera González

Editors

Radio Design in Nanometer Technologies addresses current trends and future directions in radio design for wireless applications. As radio transceiver is constituting the major bottleneck in a wireless chipset in terms of power consumption and die size, the radio must be designed in the context of the entire system, end to end. Therefore the book will address wireless systems as well as the DSP parts before it gets into coverage of radio design issues. To that end, the book contains three parts:

Part 1, a general part discussing current and future wireless networks, chipset evolution over the past decade and ending with a discussion on radio requirements for software defined radio (SDR).

Part 2 will focus on the digital baseband of a wireless chip set, flexible DSP cores for multi-standard wireless platforms and system on chip SoC implementation and design flow issues.

Part 3 will be devoted to radio design issues starting at the transceiver level and going down to discrete critical issues facing design of future multi-band multi-standard radios for emerging wireless standards such as UMTS, WiMaX, MIMO and WLAN in a way that is consistent with the prevailing vision of SDR.

As such, the book is the first volume that looks at the integrated radio design problem as a 'piece of a big puzzle', namely the entire chipset or single chip that builds an entire wireless system. This is the only way to successfully design radios to meet the stringent demands of today's increasingly complex wireless systems.

Radio Design in Nanometer Technologies

Springer

RADIO DESIGN IN NANOMETER TECHNOLOGIES

Radio Design in Nanometer Technologies

Edited by

MOHAMMED ISMAIL
The Ohio State University,
Columbus, OH, U.S.A.

and

DELIA RODRÍGUEZ DE LLERA GONZÁLEZ
The Royal Institute of Technology,
Stockholm, Sweden

*Ismail would like to dedicate
this work to his family.
Para todos los que han hecho
de mí una persona mejor.
Con todo mi cariño. Delia*

Contents

List of Figures

List of Tables

Preface

As we move beyond third generation (3G) wireless, future handheld wireless devices will be able to access different wireless infrastructures, e.g. cellular, WLAN, WiMaX for a multitude of wireless services including voice, data and multimedia applications. As a result, the radio part of a chipset for such a device will be increasingly complex and challenging. Currently, commercial radio chips are designed in 0.18 and 0.13 micron CMOS technologies. Single chip solutions (radio plus digital baseband) are recently becoming available commercially for Bluetooth, WLAN and GSM. Soon, radio chips will be designed in nanometer (<100 nanometer) technologies. This poses another significant challenge particularly that mask set costs increase exponentially with smaller feature size, market windows are getting narrower and product life cycles are becoming shorter. All this requires that fully integrated radio design achieve first-pass-silicon success.

This book addresses these challenges and discusses key aspects of integrated radio design for future handheld wireless devices. Recognizing the fact that a successful radio design must be done in the context of an end-to-end system solution, the book discusses trends at the wireless network and system levels as well as trends in programmable system-on-chip (SoC) digital baseband solutions and in programmable RF CMOS radio transceivers. To our knowledge, this is the first text on the subject of integrated nanometer radio design and the first to address the radio design problem in the context of a complete end-to-end wireless solution.

By looking at the requirements of super 3G (aka UTRAN/LTE or long Term Evolution), one can see that integrated radio systems of tomorrow will be very complex. Current and future trends call for pushing system integration to the highest levels in order to achieve low cost and low power for large volume products in the consumer and telecom markets ,such as feature-rich handheld battery-operated devices. While CMOS technology scaling to nanometer levels, coupled with innovations in platform based systems and Network-on-Chip

(SoC and NoC) have resulted in great strides with the digital part of a system, the analog, radio or mixed signal part of the total solution remains a major bottleneck. Random process variations do not scale with feature size leading to over design and increased power consumption. Lack of accurate process, package models and RF design kits presents another challenge. Therefore, in today's analog RF design environment, a fully integrated CMOS radio may require several silicon spins before it meets all product specifications and often with relatively low yields. This, results in significant increase in NRE costs, especially when mask set costs increase exponentially as feature size scales down. Furthermore, this could lead to missing important market windows, particularly with the decreasing life cycles of semiconductor products. The choice of topics covered in the book is motivated by the need to minimize integrated RF design risks and to reduce silicon spins.

The book is divided into three main parts. Part I has three chapters and deals with current and future trends in wireless communications and the evolution of wireless chipset development. Part II has four chapters devoted to digital baseband cores and their mixed signal interface to the radio. Part III has 6 chapters devoted to key aspects of fully integrated radio design.

Chapter 1 presents a futuristic view of next generation wireless networks and discusses challenges in system architectures and communication paradigms. Chapter 2 discusses cellular RF requirements and gives an overview of the evolution of cellular chip sets and of the integration trends. Chapter 3 focuses on challenges and design solutions for software defined radios.

Chapters 4 and 5 are devoted to system-on-chip (SoC) design and implementation of programmable digital baseband process cores while Chapters 6 and 7 are focusing on mixed signal and data converters to interface with the digital baseband.

Chapter 8 launches the radio design part of the book and discusses a methodology for the systematic design and optimization of integrated radio receivers. Chapters 9 and 10 discuss key RFIC design aspects of receivers and transmitters respectively while Chapter 11 discusses modeling and computer aided design of on-chip inductors. Chapter 12 deals with design challenges of frequency synthesizers in nanometer technologies. Chapter 13 concludes the book with RF design techniques that minimize design risks, avoids over design and achieves first-pass silicon success.

The book is intended for use by graduate students in electrical and computer engineering as well as system, analog/RF and digital design engineers in the semiconductor and telecom industries. It will also be useful for design managers, project leaders and individuals in marketing and business development.

This book has its roots in lectures by leading experts in the field from both industry and academia given as part of the RaMSiS (Radio and Mixed Signal Integrated Systems) Summer School on Radio Design in Nanometer Technolo-

gies held in Visby, Gotland, Sweden in the summer of 2005. We would like to thank all those who assisted us at different phases of this work specially our colleagues of the RaMSiS Group, the Swedish Royal Institute of Technology and of the Analog VLSI Lab at Ohio State. Special Thanks go to all authors for their very valuable and timely contributions, see a complete list of their names and affiliations in the acknowledgements section. We would also like to thank the Springer crew, especially Cindy Zitter for all her help.

Finally, but not least, we would like to thank our families for their understanding and support during the development of this work.

Mohammed Ismail and Delia Rodríguez de Llera González
Stockholm, Sweden
July 2006

Acknowledgment

We would like to acknowledge all contributors to this book. We list them below in an alphabetical order:

Waleed Khalil
Intel Corporation, Phoenix, Arizona, USA.

Peter Klapproth
Philips Semiconductors BV, Eindhoven, The Netherlands.

Dake Liu
Linköping University, Linköping, Sweden.

Sven Mattison
Ericsson Mobile Platforms, Lund, Sweden.
Anders Nilsson
Linköping University, Linköping, Sweden.

Delia Rodríguez de Llera González
Royal Institute of Technology, Stockholm, Sweden.

Elyse Rosenbaum
University of Illinois, Urbana, IL, USA.

Ana Rusu
Royal Institute of Technology (KTH) Stockholm, Sweden.
Technical University of Cluj-Napoca, Romania.

Henrik Sjöland
Lund University, Lund, Sweden.

Christer Svensson
Linköping University, Linköping, Sweden.

Eric Tell
Linköping University, Linköping, Sweden.

Niklas Troedsson
Lund University, Lund, Sweden.

James Wilson
Firstpass Technologies, Inc., Dublin, Ohio, USA.

Jens Zander
Wireless@KTH, Royal Institute of Technology, Stockholm, Sweden.

PART I

CURRENT AND FUTURE TRENDS

Chapter 1

"4G" AND THE WIRELESS WORLD 2015 - CHALLENGES IN SYSTEM ARCHITECTURES AND COMMUNICATION PARADIGMS

Jens Zander

1.   Introduction

Wireless applications and services are in the next decade likely to become a pervasive, with a widely spread use of wireless devices everywhere. The technology will undergo a transformation, from an expensive, highly visible, "hi-tech" technology as in early cellular phones, over the current state were (almost) everyone owns a mobile phone, to a "disappearing technology" that is present everywhere and taken for granted. Since the current cellular mobile approach, with its excellent mobility management and coverage properties, does to not scale in an economical fashion into large bandwidths, it is more likely that a highly heterogeneous infrastructure will emerge with a large variety of wireless access options. Such a vision challenges many of the current paradigms in mobile communication. In this paper we will discuss these challenges in more detail and give an outlook on some possible directions for research and developments.

Computation and wireless communication capabilities are radically integrated in a great variety of different everyday things, from simple sensors and interactive appliances (cards, rings, eyeglasses...), via pocket and lap-sized devices to wall or table screen working areas. The technology will undergo a transformation, from an expensive, highly visible, "hi-tech" technology as in early cellular phones, over the current state were (almost) everyone owns a mobile phone, to a "disappearing technology" that is present everywhere and taken for granted. The consequence of this vision is that not only wireless terminals but also infrastructure components (similar to electric appliances) need

to be no-maintenance/disposable and self-configuring, local access networks that can be deployed in minutes without requiring highly skilled and trained personnel. This will radically lower the entry thresholds for new actors in the infrastructure field which creates new business opportunities and competition. Facility, shop and restaurant owners and even private persons will provide both wireless access to global services as well as "value-added" localized services. The infrastructure components can form integrated parts of a "wireless grid" and be accessible to the public. The large diversity and efficient competition between providers of network & services elements or combinations thereof will provide seamless service according to user preferences. The user priorities tend to change from good coverage to low cost when networks have been deployed in large scale and getting mature.

Two of the key design challenges of the next decade will therefore be
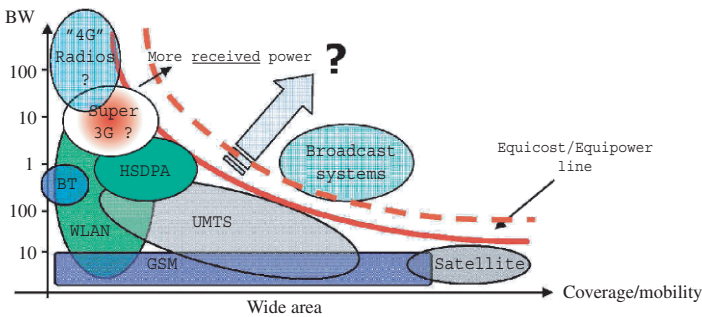
Figure 1.1. Range/Coverage/Mobility - Bandwidth relationship

- How to provide efficient services on a heterogeneous wireless access infrastructure with many network components owned by various business players providing a plethora diverse service offerings,

- How to deploy local high bandwidth access infrastructure in an highly efficient and economical way.

In the following sections we will see how these challenges can be met and what critical research issues still remain open.

2. From the "swiss army knife" ....

As wireless infrastructure system approach higher and higher data rates, the ranges of the individual radios becomes less and less as is illustrated by Figure 1.1[1]. This graph in a coarse manner illustrates the range/bandwidth relations of various wireless access systems. All systems are confined to the lower left region, bounded by the solid line, which is basically the "Shannon bound" of communication theory. This limit is due to the fact that a minimum amount of

energy is required to reliably transfer every bit of information. The higher the data rate, the less energy is available for every bit (at constant transmit power) at the transmitter, the less is the range. If we want to push this limit upwards to the right we need either increase the transmitter power[2] or to increase the number of access points. The former path seems closed due to high expectations on battery life and to limit potential health hazards. The latter path on the other hand involves larger investments. It has been shown [1] that the number of access points required grows linearly with the bandwidth provided. Besides high bandwidth, important cost drivers are reliable wide area coverage, high-speed mobility and real-time requirements. In fact, it seems impossible to provide all these four properties simultaneously for reasonable costs. It is for instance reasonably cheap to provide high data rates for internet access (non-realtime) to pedestrians (low mobility) in city centers (small coverage), whereas providing a real-time high rate service to fast mobile terminals all over a sparsely populated country is vastly more expensive.

From a business point of view, the problem is quite clear. An operator needs a sufficiently large number of users for every access point he deploys to eventually recover his investments without claiming un-acceptable prices for his access service. This means that in areas where the user density is low, only a low density of access points can be supported with a corresponding low to moderate data rate. On the other hand, in locations where the user density is high, high data rates is not a problem.

The traditional solution to this problem is to provide access systems with flexible air interfaces that can provide both high rate, short range as well as low rate long range communication. A leading paradigm behind such a solution is that the user terminal can handle only one (albeit very complex) air interface and that a single, world-wide standard is necessary for commercial success. The key drawback of such a "Swiss army knife" system (one system reasonably suited for all purposes) is its complexity and it inherent lack of flexibility. The system will be capable of reliable wide area coverage and high speed handover everywhere, but pedestrian (lap-top) users in city center will only very rarely need those capabilities. In addition there is a risk that the built-in flexibility in the system is not sufficient to meet future needs. Wide-area infrastructure deployment is a matter of decades, whereas user needs may change more rapidly. The large investments required to provide single system coverage everywhere is also a significant barrier against new entrants on the market and effective competition.

3.   ... to Navigating the "Wireless Chaos"

A much more attractive scenario would be triggered by the appearance of a flexible multimode terminal. In this scenario, a plethora of specialized access systems would co-exist, each optimized to provide cost effective access for its

"niche" market (geographical, mobility, Quality-of-Service etc) without any requirement (on each and every system) for coverage and service everywhere. This would indeed challenge that a single world wide wireless standard is necessary, that a single public operator is needed to provide cost effective physical access. On the contrary one could claim that for broadband local access systems, very high data rates are not that difficult to achieve in a cost-efficient manner provided that high speed mobility may not be the prime interest of the users. In local wireless access the boundaries between fixed and wireless become blurred and local access provisioning will become more and more the business of facilities owners. In rural areas and for vehicular mobility on the other hand, the traditional cellular solutions are the most cost effective ones and are likely to keep their dominating position.
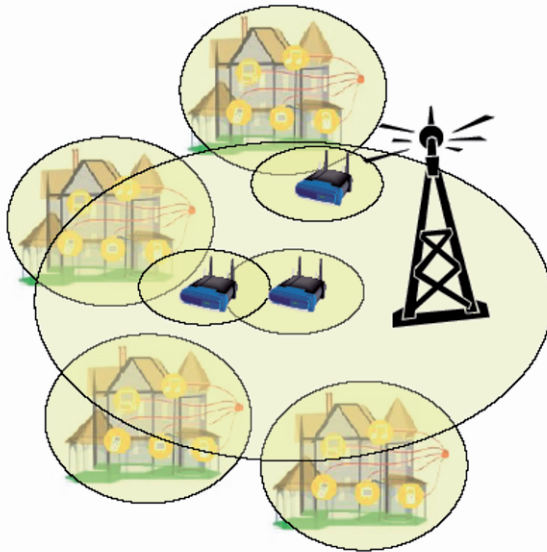
Figure 1.2. Heterogeneous wireless infrastructure with a multitude of access with varying properties (range, data rate, mobility etc.)

In this mixed environment, infrastructure deployment can be incremental and the entry thresholds for new niche actors are significantly lower than today. Infrastructure components (access points, routers etc) have to become low cost, user deployable, i.e. self-configuring and low maintenance [3]. In this scenario it is not the individual access schemes, but instead the collection of systems that will provide the universal coverage that we aspire. Terminals are consumer products with a life-cycle of 2-3 years and they are perceived by the buyer to be intimately connected with the applications of choice. Choosing the proper

set of air interfaces for a particular terminal is thus not a problem and when each he buys a new terminal it is likely to contain a different set. Still the user wants access to his services in a transparent way and he is likely to be ignorant regarding which air interface his terminal is currently using. The economic advantages of this scenario are significant to all actors - operators do not need to provide coverage everywhere and customers will benefit from effective access competition.

The strict compatibility at the physical layer of the current systems will thus be replaced by interoperability at the network layer. Creating this interoperability in future access is is addressed in the IST/FP6 project WWI Ambient Networks [3] which proposes a modular "generalized" internetworking approach, where the provisioning of connections can be achieve across technology and business barriers. The latter is imperative if a competitive environment is to be achieved in the wireless access domain to the benefit of the consumer.

Whereas a key paradigm in traditional cellular system was the spectrum is scarce resource, this no longer holds in our scenario. An important consequence of relaxing the physical layer compatibility, is that different access systems can use different parts of the spectrum. Further, as we go to higher data rates (shorter ranges), spectrum reuse become more effective and international spectrum coordination becomes less of a problem as (short range) signals are less likely to cross national borders. More capable radios also open the possibility of a more effective use of the spectrum - not even the access system of a single provider needs to use the same part of the spectrum in every geographical area and the system may "scavenge" for free spectrum, so called Dynamic Spectrum Access (DSA) [6].

4.   Six "Grand Challenges" in Wireless Systems

To make a scenario as outlined above possible we believe that the following six research challenges have to be adequately met. These challenges are based on the work in [9]:

I. Scalability and affordability - Creating a wireless communication infrastructure for affordable, mass-market services

As the cost of providing advanced wireless devices continues to decrease, designing cost effective infrastructure solutions capable of providing affordable wireless broadband access (almost) everywhere is one of the key success factors for future wireless systems. This research challenge includes devising novel radio technologies, new system architectural concepts, and new and cost-efficient ways to provide attractive services to end-users.

II. Seamlessness and Transparency - Providing services independently of system technology

One of the success factors of IP networks is the end-to-end principle, which separates services and applications from bit transport. The same service can

be provided on a variety of devices (using higher level protocols) without any change in the infrastructure. A key challenge is to preserve such architecture in order to enable easy and dynamic composition of disparate networks amid an ever-increasing heterogeneity of technologies and infrastructures. An additional difficulty is to provide access and services across networks operated by different business actors from various sectors, such as telecommunications, automotive, transportation, medical, industrial control systems etc.

III. Mastering complexity of interaction - Providing high quality services on the edge of technology and artefacts that are easy to use for everyone

A great challenge is how to provide an easy to use, natural, stable, and convenient interface to the user in each situation, in spite of the great complexity of the underlying system. This involves personalized human interfaces, understanding a complex interplay of behaviors, as well as adaptivity and context sensitivity of services and applications.

IV. Zero-configuration and reliability through massive redundancy and network robustness - Lowering entry thresholds for new actors in the wireless system market by low cost, simple-to-deploy, and low-maintenance systems and networking components

Future wireless devices and infrastructure components have to be deployed and maintained by owners or users without specific skills and special training. This means that the devices need to be adaptive and self-configuring, sensing their physical and logical environment. The key challenge is to exploit massive redundancy and adaptivity to build secure, robust and highly reliable networks and systems from large number of consumer grade devices.

V. Regulative environment - Lowering regulatory entry barriers for new actors to stimulate the innovation process

The continuous process of international allocation of frequency bands has normally a delivery time of more than ten years. Poor utilization of the frequency spectrum as well as high entry barriers for new products developed by e.g. SMEs may have a detrimental effect on the innovation system. Research must include exploration of new radio technologies and regulatory regimes that allow for a more dynamic frequency sharing. A key research challenge is non-cooperative inter-networking and radio resource management.

VI. Policies and Business models - Economic feasibility of new technologies and architectures

To ensure commercial viability we must identify the business roles and interfaces as well as deployment concepts. New business scenarios have to be developed. These must allow different size and types of players to compete and cooperate, thus enabling new business models based on established trust relationships. The choice of technologies and system architectures heavily depends on these business and policy models [7][8].

5.   Challenges in Radio Design - Flexible or Software Defined Radios

What are the consequence of the scenario with respect the work on radio design which is the key topic of this book? We can identify two key properties of terminals operating in the scenario described above:

5.1   Multi-mode Capability

Future terminals need to be capable of switching between several air-interfaces literally on the fly. This can be done either by integrating separate hardware for each access mode in the terminal or by using programmable hardware, that potentially could reuse the same hardware and have the various access modes defined in software (Software Defined Radios). The latter would have the interesting property that the terminals could be reconfigured during operation (after they left the factory). It is however questionable, if this property is really useful since the expected life cycle time in terminals is short compared to the deployment rate of new infrastructure (and thus the appearance of new air-interface standards). Terminals are more likely to be tailored to user needs and specific application and than disposed of, rather than recommissioned for some other purpose. A good example is the laptop PC: the vast majority of users use pre-installed applications and never install new software on their own. Reprogramming is limited to maintenance and updates of existing software, which very rarely goes to the hardware level. A more reasonable approach for SDR use is therefore to use "Firmware Defined Radios" were programming radios is more a matter of efficient production of terminals rather than a tool for "on-the-fly" flexibility. Furthermore true flexibility ("future proofness") requires significant performance margins for which we pay in power consumption.

5.2   Spectrum Agility

The Multimode terminals will need to operate over large frequency ranges in order to facilitate a more dynamic spectrum management. The focus is here on wide frequency ranges rather then reconfigurability and programmability. In this respect the radios do not need to be software defined nor in it self "cognitive". The latter term is currently frequently used to describe a radio a frequency agile radio system in combination with adaptive scheme for frequency management. Also a "cognitive" radio system does not need to be software defined.

6.   Conclusions

In this chapter we have briefly outlined some important trends in wireless systems and how these are driven by strong economic factors. The proposed heterogeneous network scenario with many co-existing standards, each tailored to its specific niche, is a direct consequence of these factors. We have

demonstrated that in future systems more functionality and flexibility will be required in the terminals, since large scale wireless infrastructures inherently cannot adapt quickly to new demands and services. Finally at the physical level, we identified two key requirements on future radios - multimode capabilities and spectrum agilities.

Notes

1   Figure 1.1 is based on the popular graph devised in the 1990s by Mike Calendar of the ITU used by the IMT-2000 planning and standardization work

2   In fact a promising approach is to increase the received power by using "smart" directional antennas making sure that the transmitted power in focused on the receiver. Large gains have so far been difficult to achieve in practice

References

[1]   Zander, J, *On the Cost Structure of Future Wideband Wireless Access*,IEEE Veh Tech Conf, VTC '97, Phoenix, AZ, May 5-7, 1997.

[2]   Zander, J, Markendahl, J,*Low Cost Broadband Wireless Access - Key Research Problems and Business Scenarios*, Invited presentation, Int Symp. Adv Radio Tech ISART2004, Boulder, Co, March 2004.

[3]   WWI Ambient Networks http://www.ambient-networks.org/

[4]   G.P. Kouduridis, P Karlsson, J Lundsjo, A Bria, M Berg, L Jorguseski, Fo Meago, R Aguero, JSachs, R Karimi, *Multi-Radio Access in Ambient Networks, "Multi-Radio Access in Ambient Networks"*, Everest Workshop, Barcelona, Spain, Nov 2005.

[5]   C Cedervall, P Karlsson, M Prytz, J Hultell, J Markendahl, A Bria, O Rietkerk, I Karla, *Initial findings on business roles, relations and cost savings enabled by Multi-Radio Access Architecture in Ambient Networks*, Proc WWRF 14, San Diego, CA, July 2005.

[6]   Dynamic Spectrum Access: http://www.wireless.kth.se/projects/DSA/

[7]   B Thorngren, J Markendahl, *Business models and business roles for use of low cost infrastructure solutions*, RVK 2005, Linkoping, June 2005.

[8]   Novel Access Provisioning: http://www.wireless.kth.se/projects/NAP/

[9]   Karlson et al., *Wireless Foresight, Scenarios of the Mobile Worlds in 2015*, Wiley 2003.

Chapter 2

CELLULAR RF REQUIREMENTS AND INTEGRATION TRENDS

Sven Mattisson

1.   Handset Technology Drivers

The cellular handset has gone through a rapid evolution. From being an executive's attache-sized cell-phone toy to end up in everybody's pocket, or from $3\,\mathrm{kg} \Rightarrow \lesssim 100\,\mathrm{g}$, $\$\,3000 \Rightarrow \lesssim \$\,100$, and from useless $\Rightarrow$ weeks of stand-by time.

In this chapter we will outline some of the challenges and trends in the handsets technology business.

Needless to say, it is the ever increasing integration capabilities predicted by Moore's law that has facilitated the rapid adoption of cellular phones. With annular sales exceeding 500 million units the volumes are high enough to exploit the most advanced integrated circuit technologies, and, in fact, the handset business has driven the development of low-power and RF technologies. With such a large market, many players are interested and competition is fierce. In the beginning phones were competing with size as well as talk and standby time, but today these parameters are mostly "good enough" and it is with the versatility of the handset, for example as personal information managers, music players, games, or cameras that manufacturers compete. In addition to the growing user-application suite, the cellular evolution with more frequency bands and cellular standards is constantly challenging the designers, as all of this has to be added without raising the manufacturing cost. Of course handsets supporting

11
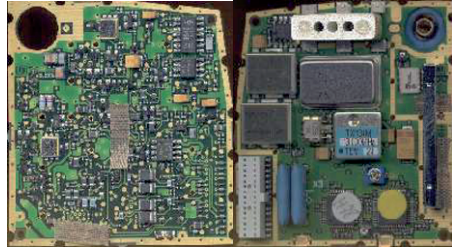
- 2 boards (←logic and radio↑)

- RF board 5000 mm^2

- 350 RF components

- 2 ASICs

- 2 standard ICs

- 12 modules (VCO, filters, PA)

- 5 V

Figure 2.1.   GH197 (1992).

new standards must also achieve similar use time and size as the more mature 2G handsets.

How has this evolution been made possible and how can we continue to deliver increased functionality at lower cost and size?

1.1   Handset Complexity

In figure 2.1 the PCB of one of the very first pocket sized GSM handsets is shown, the Ericsson GH192 from 1992. This phone has a single band (900 MHz) digital radio modem[1] and occupies two printed-circuit boards (PCB). Some seven years later, the Ericsson T28, see figure 2.2, supported two bands (900 and 1800 MHz) on a single PCB with the RF part only some 25 % in size of that of the GH192. The next step, see figure 2.3, which was launched circa a year after the T28, now has added a third frequency (1900 MHz), enabling roaming between, for example, Europe and the USA, with an even smaller RF board area. To summarize the changes when moving from the GH192 to the T39 we see no reduction in the number of application specific integrated circuits (ASIC) but a reduction from $2 \rightarrow 1$ in standard integrated circuits (IC), $12 \rightarrow 5$ in modules, $350 \rightarrow 90$ in RF board components, while, the number of RF bands increased from $1 \rightarrow 3$, clearly indicating that much functionality has been moved from
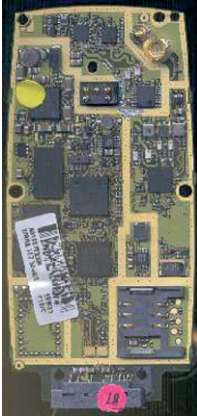
- dual band

- 1 board, single sided mounting

- RF part 1300 mm^2

- 140 RF components

- 3 ASICs (RF, PA, TX/VCO)

- 1 standard ICs

- 4 modules (X-tal, filters, antenna switch)

- 3.6 V (2.7 V)

Figure 2.2. T28 (1999).

the PCB components and modules into the ASICs resulting in a five-fold PCB area reduction in-spite the increased RF band support. These handset examples clearly show that a chip-count reduction and integration strategy was used to push size and cost down while increasing functionality.

1.2   Chip-count Reduction Strategy

In the examples in figures 2.1–2.3 the radio architecture went from a super-heterodyne receiver with an offset-loop transmitter, see section 3, requiring several external components, to a homodyne receiver with a direct phase-modulation transmitter. The GH197 used an external receive (RX) channel filter and a transmitter (TX) image filter in addition to the band-select filter between the antenna and the low-noise amplifier (LNA). In the T39 only the band-select filters (now three because of the triple-band operation) are external to the transceiver ASIC. Furthermore, the GH197 employed two transceiver ASICs in addition to a VCO module and an off-the-shelf synthesizer IC, while the T39 has all these functions integrated in the transceiver and base-band ASICs. The number of RF components also emphasize the benefits in integrating as much as possible into the transceiver ASIC.

- triple band

- 1 board, single sided mounting

- RF part $1000 \, \text{mm}^2$

- 90 RF components

- 2 ASICs (RF, PA)

- 1 standard ICs

- 5 modules (X-tal, filters, antenna switch)

- 2.7 V

Figure 2.3.   T39 (2000).

Today, a typical single-mode multi-band handset has one RF transceiver ASIC, (with external power amplifier, band-select filters, antenna switch, and crystal or TCXO[2] module), one digital baseband ASIC, and, one mixed-signal power management ASIC (possibly also with some audio functionality) to cover the basic radio-modem and user-interface (e.g. display driver, phone book etc.) functions. The reason for this ASIC partitioning is that no ASIC technology node has been able offer a competitive solution for all three parts on a single die.

The main reasons for the lack of competitiveness of the single-chip (or system-on-a-chip, SOC) solution has been the rapid cellular and application evolution requiring frequent updates of the baseband ASIC, while the update-rate of the RF and mixed-signal ASICs is more relaxed; see figure 2.4 for a tentative cellular modem block diagram. When all parts are implemented on the same chip, an update of any part requires the others to be updated as well [1]. Furthermore, the baseband has to move to new technology nodes faster than the other ASICs for cost reasons. In fact the RF and mixed-signal parts often use more mature technology nodes than the digital base-band does, and are, thus, 2–5 technology nodes (feature-size-vise) behind the digital ASIC. Furthermore, the RF and mixed signal designs have additional requirements (e.g. accurate
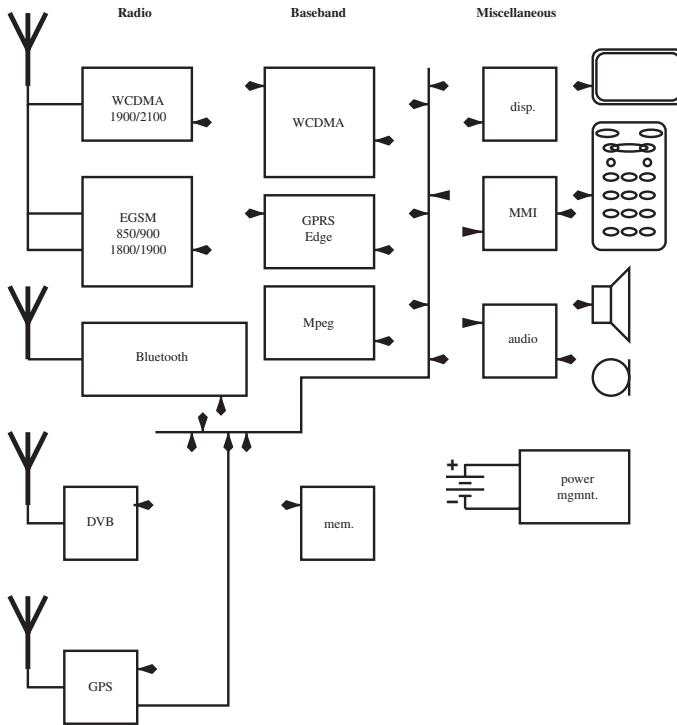
Figure 2.4. Tentative handset block diagram.

RF modeling, high-performance passive components, and, high-voltage capability) and to add such features to the digital technology will delay process development leading to larger chip areas for the digital parts compared to an implementation in the most recent digital(-only) technology node. The individual chip complexities and sizes have also been large enough to make the risk of a SOC implementation of a cellular hand-set modem too high given the potential cost advantage.

Multi-chip modules (or system-in-a-package, SIP) is an alternative to the single-chip ASIC, offering similar advantages in terms of size and ease-of-use for the handset maker, but without the problem of relying on one ASIC technology only for all the circuits [2]. However, multi-chip modules have been too expensive and have had some reliability issues in the past delaying their commercial introduction in cellular handsets.

The question is now how to continue the handset integration when the chip-count reduction strategy seems to be close to roads end with three incompatible ASICs. It is clear that a straight reuse of legacy circuit techniques will not provide an answer. New architectures and trade-offs will be needed [3]. Also, the ASIC and module technology evolution changes the rules such that build-
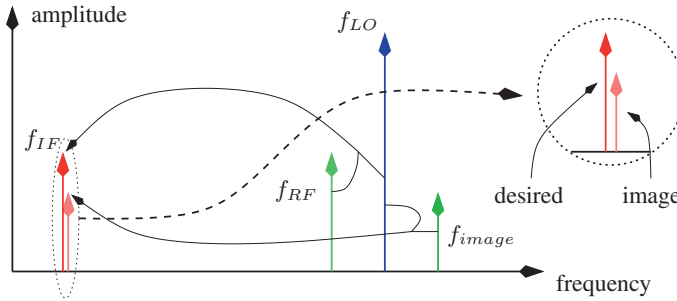
Figure 2.5.   Image response.

ing techniques that were not competitive in the past may suddenly be worth considering again. For example, single-chip modems for Bluetooth has been announced since some time (e.g. [4, 5][6]) as well as modules, and that begs the question when will this be competitive for cellular handsets and which technology, SOC, SIP, or, a combination of both. To address this question we will look into the design challenges of a cellular modem, or transceiver.

2.   RF Transceiver Design Challenges

The typical task of an RF transceiver is to convert an incoming radio signal to a digital bit stream in the RX chain and to modulate a radio carrier with a digital bit stream in the TX chain. All these operations are not necessarily done in one SOC or SIP, but we will discuss that later. Todays RX chains almost always include band-select filters, low-noise amplifiers, frequency translation, (coarse) channel-select filtering, analog-to-digital conversion, and, finally digital decimation and channel select filtering. The TX chain contains the inverse functionality.

Depending on the details of the cellular standard and frequency band, the over-all requirements vary, but also the choice of radio architecture for the RX and TX chains has a major impact on the detailed requirements. In the following we will, for simplicity, focus on the receiver and highlight the various implementation issues and how they can be solved.

2.1   Frequency Conversion

Frequency translation, to or from an intermediate frequency (IF), is accomplished by multiplying the information signal with a local oscillator (LO) signal. Traditionally this has been done in a switching mixer (see for example [7, chapter 12]) where the LO signal is a symmetric square wave. A periodic signal can be approximated with a sum of sinusoids, but for simplicity let's consider only the fundamental tones. When these tones are multiplied we get a sum and

a difference tone. Because of this trigonometric relationship any mixer will be sensitive to both the sum and difference of the RF and LO signals, see figure 2.5. A down-conversion mixer will output the difference (the sum is generated but filtered out) and because two RF frequencies ($f_{RF_1} - f_{LO}$ and $f_{LO} - f_{RF_2}$ both will generate the same difference frequency, a mixer is responsive to both RF signals, the wanted signal and its image.

Two common techniques exist to suppress the image frequency: filtering and quadrature mixing, see e.g. [7]. In the first case an image-reject filter is put in front of the mixer to suppress the image sufficiently to render it harmless when it superimposes on the wanted signal. Since the image frequency is spaced two IF frequencies away from the wanted RF signal, and the IF is a low frequency, the image filter has to be very steep (i.e. selective). For stability and noise reasons such a filter is often made out of passive components, e.g. a surface-acoustic wave (SAW) filter, which is incompatible with present ASIC technology. Thus, to increase the level of integration quadrature mixers are commonly employed to eliminate the need or to relax image-reject filter requirements. Since a quadrature mixer can be implemented by two regular mixers and a phase shifting network, this can easily be integrated on an ASIC.

Recently a lot of attention has been put on sampling as a frequency translation technique. Sampling, however, performs the same basic function as the mixer, except that the LO waveform now may be an impulse train or something similar. In some aspects sampling and mixing differ but they share the same fundamental issues. The image response is, for example, similar to folding distortion in the sampler. Similarly sampling jitter and LO phase-noise are just two different ways of describing noise transfered from the LO to the IF signal.

2.2   Noise

Thermal noise (i.e. white noise) is ultimately setting the limit on how weak a signal that can be properly decoded. The available noise power density is $n_0 = kT$, where k is Boltzmann's constant and T absolute temperature, and the noise power in a certain bandwidth is $n_0 B$, where B is the noise bandwidth (see e.g. [8]). For a properly designed cellular receiver with discrete-time baseband signals, the noise bandwidth equals the inverse of the symbol rate [9]. Thus, as long as no noise folding occurs in the sampler, the analog (i.e. continuous-time) filters have no direct influence on the receiver noise bandwidth. Of course the analog filter shape influences inter-symbol interference (i.e. group-delay ripple [8]) and suppression of off-channel interference, and may , thus, indirectly impair the receiver performance.

In addition to white noise, electronics also have colored noise like 1/f-noise at low frequencies. At higher frequencies the noise power density is proportional to f^2, but this is typically white noise that increases due to gain roll-off with frequency.
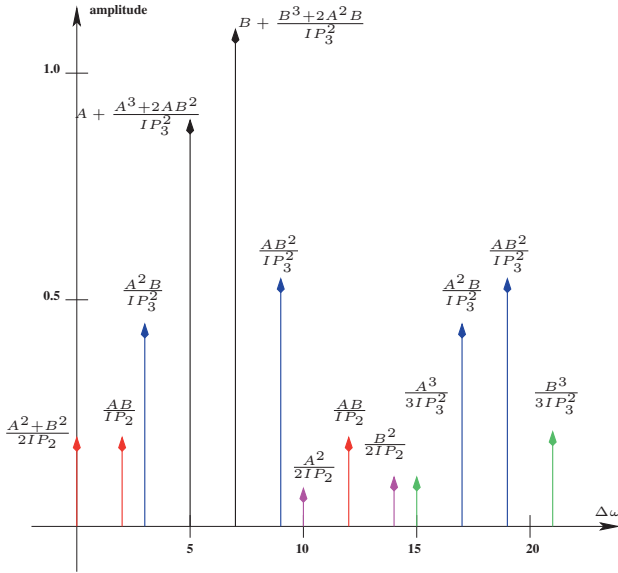
Figure 2.6.   Intermodulation products resulting from $0.9\cos(5\Delta\omega t) + 1.1\cos(7\Delta\omega t)$ applied to a cubic nonlinearity with $IP2 = 5$ and $IP3 = \sqrt{2}$.

It can be shown, see for example [7], that a transistor has a output noise current density $i_n^2 \simeq 4kTg_m$, where g_m is the device transconductance, that is, it is roughly as noisy as a resistor when $R = 1/g_m$. This can be used as a rule-of-thumb for bipolar and MOS transistors as well as diodes. Furthermore, it can also be shown that a capacitor C has an equivalent[3] noise voltage density $v_n^2 = kT/C$. For a given filter time constant there, thus, exists a direct relation between noise floor, current consumption, and, area as $\tau = RC \simeq C/g_m$ and $g_m \propto I_{DS}$.

2.3   Linearity

All electronic components are more or less nonlinear. That is, above some signal level, the signal superposition principle is not a valid approximation any longer. Assuming the device is only weakly nonlinear we can approximate its transfer characteristic (e.g. gain) by a cubic polynomial like $y = f(x) = a_1 x + a_2 x^2 + a_3 x^3$ (see e.g. [10]). Assuming that the input signal is a sine wave, harmonic tones at integer multiples of the input frequency, that is harmonic distortion, will be generated. With two sinusoids as input we get their harmonics as well as intermodulation tones at frequencies which are linear combinations of the inputs and their harmonics, i.e. $f_{IM} = \pm M \cdot f_1 \pm N \cdot f_2$, where M, N are positive integers, see figure 2.6. The order of the intermodulation product equals $N + M$ and it is common to describe the linearity of a device, or a circuit,
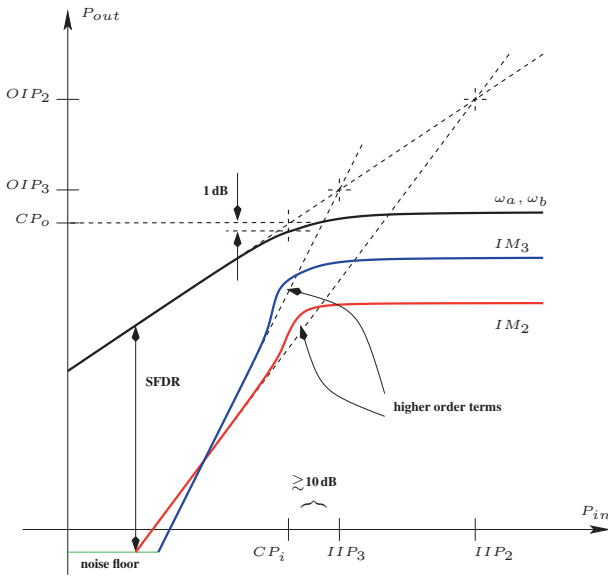
Figure 2.7. Linearity definitions; intercept points, compression point and spurious-free dynamic range.

by the intercept point, where the (extrapolated) intermodulation product equals the inputs in amplitude, see figure 2.7. For a cubic nonlinearity we then get $IP_2 = |a_1/a_2|$ and $IP_3 = \sqrt{|4/3 \cdot a_1/a_3|}$ for the second- and third-order intercept points, respectively.

When the input, or output, signal level approaches a corresponding intercept point, the distortion products associated with that nonlinear coefficient will become significant. Of course, what is significant depends on the situation at hand. For example, gain can be shown to deviate by 1 dB from its linear value when the input level is approximately $IP_3 - 10$ dB, and this is referred to as the gain compression point, CP. The compression point for a bipolar transistor is similar to its quiescent current or to the maximum voltage swing, whichever is reached first. The dynamic range of a radio can then be defined as the distance (in dB) between CP and the noise floor or alternatively as the spurious-free dynamic range (SFDR), see figure 2.7, which is defined as $2/3 \cdot (IP_3 - N)$ (in dB) where N is the receiver noise floor.

Some odd-order intermodulation products, between in-band signals, show up close to the in-band signals and will then also be in-band signals. These are typically governed by the IP_3 characteristics which is the primary linearity parameter. For example, two adjacent channels will cause an on-channel intermodulation signal that will add to the noise inside the channel bandwidth. In this case, signal levels far below IP_3 may generate enough intermodulation
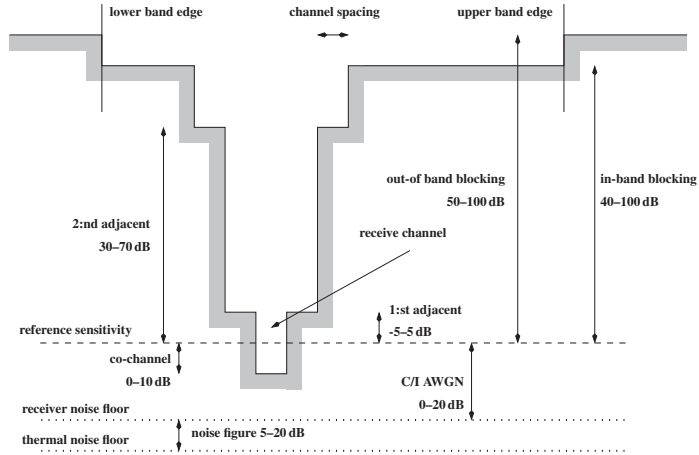
Figure 2.8.   Selectivity and blocking dynamic range requirements.

noise to desensitize the radio. Second-order distortion products show up at sum and difference frequencies and will be harmful first when the wanted signals are converted to a low IF frequency.

2.4   Selectivity

A radio receivers ability to receive a certain signal (i.e. channel) is given by its selectivity, see figure 2.8.

On the desired channel, interference and noise is called co-channel interference. Depending on the modulation type this co-channel interference must typically be 0–10 dB below the desired signal. To enhance the co-channel performance the detector typically has to be improved, filtering is less effective as it also affects the wanted signal.

Either side of the wanted channel we have adjacent channels. The closest adjacent channel is normally not suppressed very much but the 2:nd, and more distant, adjacent channels has to be suppressed anywhere from 30–70 dB. This is typically accomplished by the combined filtering of the analog and digital IF filters. The main task of the analog filter is to limit the required dynamic range of the analog-to-digital converter (ADC) and to prevent noise folding into the signal sample. The digital IF filter then provides close-in selectivity filtering and group-delay equalization (the on-channel noise and interference into the detector should ideally have a white noise characteristic, including radio channel artifacts [9]). Since the adjacent-channel filters have to be very steep this task is best accomplished at a low frequency when the relative filter steepness is minimized. Today, it is common to integrate the analog filter on the
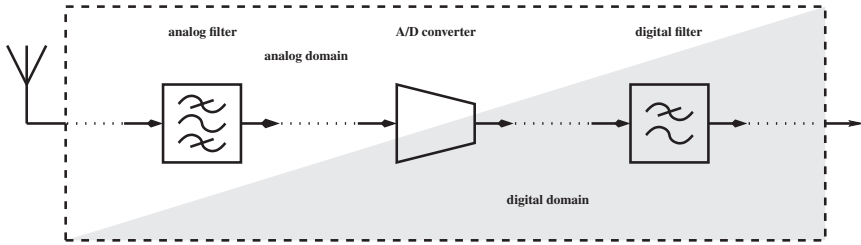
Figure 2.9. Receiver domain partitioning.

transceiver ASIC as a low-pass filter, which only requires resistors, capacitors, and, amplifiers.

Out-of-band signals have to be further suppressed and this is typically accomplished with filters at RF frequencies. Because of the high center frequency these filters have a very narrow relative bandwidth and are implemented, for example, in SAW technology and cannot be integrated. Without such filters the receiver would need a dynamic range (compression point to noise floor) around 120 dB for GSM. For systems like WCDMA, where the receiver and transmitter operate simultaneously, the band-select filter is accomplished by a duplexer, which is also suppresses the transmitter by some 50 dB (as seen from the receiver).

2.5 Domain and ASIC Technology Choices

The various building blocks of the cellular radio transceiver has to be implemented with many different goals in mind. Low power is key; todays GSM[4] cell phones consume some 2 mA in stand-by mode, which translates to some two weeks of stand-by time[5]. Low cost is another must; competition is fierce and cost is one important factor. Cost is proportional to circuit size and dependent on technology choice. Flexibility is good as it will allow for some tweaking when performance is at risk or the specifications are modified. Generic solutions and reconfigurability simplify migration and reuse but may have some impact on the circuits size. All these things must be balanced to find a solution that is robust and designable with the available resources, while supporting several frequency bands and cellular standards.

Analog implementations offer operation closer to the process limits while digital offer more flexibility and robustness. As process technology evolution provides more speed and smaller area this favors a migration towards digital implementation, see figure 2.9. The ADC is, however, a key component and a "digital solution" is more demanding on the ADC, than a more analog one, because of the increased dynamic range and sampling rate required when less analog filtering is used.

For the digital baseband there really is only one technology choice: CMOS. Still, there are many different trade-offs to be made. How should the memory be partitioned, off-chip or on-chip, should non-volatile on-chip memory be used and how to provide flexibility in the memory foot print? How to provide flexibility is not a problem as such, but flexibility has a direct impact on the cost and the challenge is to find the right match of features, flexibility and cost. Presently most memory is off-chip because this is a convenient way of allowing flexibility, as much of the feature-set expansion is accomplished via SW rather than HW changes and a low-tier phone can then use the same ASICs as a higher-tier phone but save on the memory size/cost.

Because the digital baseband ASIC typically is much larger than the other ASICs it usually pays off to go with the latest available technology node and to migrate roughly on a yearly basis (i.e. to track Moore's law). This is not the case for the RF as the chip size is much smaller and because RF design requires accurate device and parasitic modeling which typically delays the RF design kit compared to the digital, even when no process modifications are done for the RF. The RF part also benefit from passive components like high-Q inductors and varicaps which are not normally available in a digital baseline process design kits. Because of passive component quality and design flexibility RF designers have favored BiCMOS technologies in the past. Because the CMOS part of a BiCMOS technology is based on an existing CMOS technology (with or without modifications) the BiCMOS technology will lag the corresponding CMOS node and also incur extra process development work. The net effect is that the digital baseband cannot be implemented in BiCMOS for size reasons and the BiCMOS technology will be more expensive due to extra process steps. For RF alone this has not been prohibitive as the RF re-design cycles are not as aggressive as the digital with a longer life time for the RF ASIC.

The main advantage in putting the RF and baseband on the same die (SOC) is a slightly lower cost; mainly because of a reduced silicon area, as fewer I/O pins, and hence fewer bulky pads and pad drivers, are needed. Another SOC benefit is that the interface between the RF and baseband can be much more flexible as more wires can be used and also since signals are on-chip only with much lower capacitive load. A single package solution in a small package size can, however, be accomplished with a SIP as well, so there is no difference in these aspects. The SOC struggles with the problem of different RF and baseband update rates, resulting in more RF redesigns (or at least the porting of an existing design to a new technology node when it is not needed from an RF perspective) but also with parametric yield issues. Digital circuits typically only suffer from defect density incurred yield losses, whereas analog circuits also have parametric losses due to random variations in device parameters. These parametric losses will cause good baseband dies to be wasted because of RF
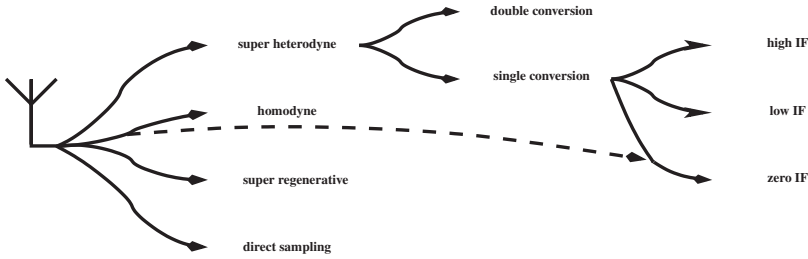
Figure 2.10. Receiver architecture genealogy.

yield which will be very bad for the over-all cost. Thus, SOC is only possible when the total RF yield is similar to defect-density limited yield.

Todays trend is to move the RF design away from BiCMOS, whenever possible, to avoid the cost premium of BiCMOS. However, not to an SOC but to an RFCMOS ASIC to decouple the baseband and RF design cycles somewhat. The cost savings in using RFCMOS highly depends on the designer being able to meet performance requirements in a similar area as in BiCMOS. Typical RFCMOS designs rely to a much larger extent on inductors (to reduce the impact of the higher drain parasitics, compared to the bipolar collector parasitics) which results in somewhat larger chip areas. Thus, the main driving forces towards RFCMOS is the fact that the BiCMOS technology development is slowing down compared to that of CMOS and to have a design in CMOS makes it more future proof. Also more digital circuitry is put in the RF die for automatic calibration and tuning as well as signal processing, which is favored by the smaller digital feature sizes in CMOS.

3.   Architectures

The radio architecture evolution reflects the component technology evolution. In the early days active components were expensive and bulky and matching was difficult. This prompted the development of the super heterodyne receiver, see figure 2.10 and [7].

3.1   Receivers

At first the super heterodyne used a single conversion stage (i.e. one mixer) but to get better selectivity (and image suppression) the double super heterodyne evolved, see figure 2.11. The double super uses a first high IF to simplify image suppression. A high IF relaxes the IF filter selectivity requirements but provides no channel selectivity so a second mixer is used to convert to a low second IF where channel select filters are more conveniently implemented.
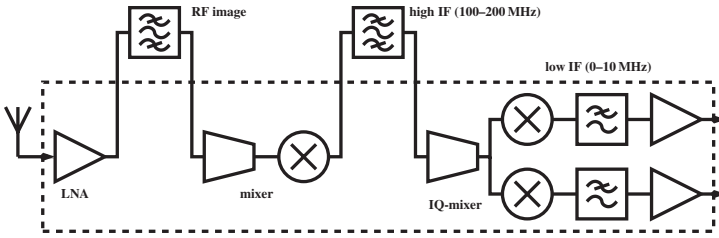
Figure 2.11.   Double super heterodyne receiver (less antenna filters).

This architecture has been the choice architecture for cellular phones until circa 1999 when the GSM homodyne appeared. The double super is very robust and is insensitive to second-order nonlinearities. The later generation double supers used a zero-frequency second IF. This choice of second IF simplifies the channel-select filters to the extent that they could be integrated. At the same time, though, second-order nonlinearities will fold stronger interference to zero-frequency and, hence, resulting in co-channel interference or noise. Because of the strong filtering performed by the image-reject and first-IF filters this sensitivity to rectification of interference was never a practical problem. Even though the channel filters could be integrated the, other two receiver filters (and the antenna filter) were not possible to integrate directly[6] and as a consequence an alternative single-conversion architecture was desirable.
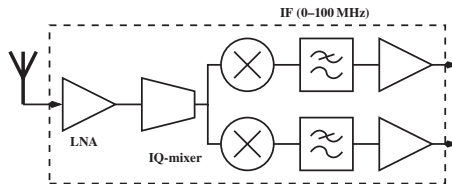
Figure 2.12.   Single conversion zero-IF receiver.

Single-conversion receivers, see figure 2.12, with a high IF were used in cordless phone applications where selectivity requirements were relaxed, but for cellular applications a low- or zero-IF receiver was required to achieve enough selectivity. The trend with wider channels (e.g. GSM with a 200 kHz channel raster) and good ASIC matching properties made the homodyne, or zero-IF, receiver feasible. The first GSM homodynes did not work satisfactorily with frequency hopping (i.e. when the radio alters the receive channel frequently to randomize interference) due to large DC offsets, and when subject to strong interference due to folding of the interference and due to cross-modulation of the interference onto the local oscillator leakage [11], both adding to the co-channel interference. With the T28, for example, the homodyne met the stricter GSM phase-2 specifications, while only requiring the antenna filter to
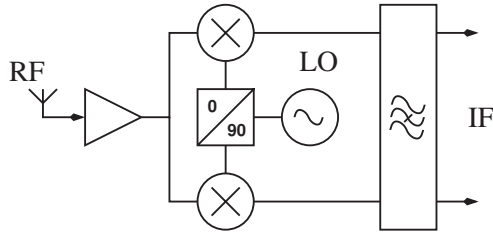
Figure 2.13.   Single-conversion low-IF receiver with poly-phase IF filter.

be off-chip, and today the homodyne or the low-IF (i.e. when the IF is around half a channel such that the IF occupies roughly DC to the double-sided signal bandwidth) is the choice GSM architecture. The first homodynes used some off-chip components (e.g. the VCO and synthesizer) but todays GSM ASICs integrate everything but the antenna filter and switches, the PA, and the crystal (or TCXO module) in the transceiver ASIC.

The homodyne, or zero-IF, receiver has the same dependence on odd-order nonlinearities as the double super but in addition also a sensitivity to even-order nonlinearities. A typical GSM receiver IP_3 requirement is some -15 dBm referred to the antenna. A switching interferer at -30 dBm must not impair the receiver reference sensitivity by more than 3 dB resulting in an IP_2 requirement around 45 dBm, or some 50 V, at the antenna. Such high IP_2 levels can only be achieved with balanced IF circuitry with very low matching errors (but this is a fundamental strength of ASIC technologies). With a homodyne, also the IF DC-offset is superimposed on the signal, and may in fact be much larger than the signal in weak-signal conditions. Also, low-frequency noise, like 1/f-noise, will add to the co-channel interference. In WCDMA, for example, when the RX signal is continuous it is possible to insert a high-pass filter in the IF path to suppress the DC offset and 1/f-noise, since the signal bandwidth is relatively large (5 MHz channel raster). This is not the case for GSM, where the settling time of any high-pass filter is too long or too much signal energy will be filtered out, and this has prompted the development of the low-IF architecture, see figure 2.13.

The low-IF receiver is almost like a homodyne but the IF is selected to be around half the channel bandwidth. This choice of IF moves the DC offset to the channel edge but it still overlaps the signal. A higher IF would solve the DC offset but then the image suppression of the quadrature mixer would not be high enough for the selectivity requirements. Hence, a low-IF receiver still has to have a very tight DC offset budget but do benefit from lower impact from 1/f-noise.
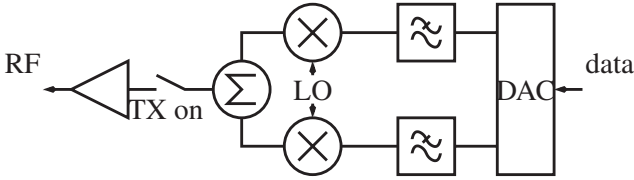
Figure 2.14.   Direct up-conversion transmitter.

To provide high data rates the trend is to widen the channel raster with each new cellular standard. The first analog systems (e.g. AMPS and NMT) used a 25 kHz raster, GSM a 200 kHz, and, now WCDMA 5 MHz. The successor of WCDMA will probably provide a more flexible choice in raster but the net effect is that 1/f-noise becomes less of a problem and the homodyne, and possibly low-IF, seems to be the receiver architecture of choice for the coming generations of cellular receivers.

In figure 2.10 the super-regenerative and direct-sampling architectures are included. The super-regenerative receiver provides very good sensitivity but poor selectivity, see for example [7], and is not a likely candidate for cellular applications (although it is, e.g., frequently used in garage-door openers). The direct-sampling receiver has received a lot of attention recently but, since sampling and mixing are both frequency conversion techniques, see section 2.1, this receiver type belongs to the single-conversion family. Software-defined radio (SDR) is another popular term but this refers to a flexible radio capable of receiving multiple-bands and -standards by having flexible, or reconfigurable, LO, ADC, and, channel-select filters. In fact, the challenge of the SDR architecture is for the analog and RF designer to cover more bands and to reuse the same filters and ADCs for varying signal bandwidths and dynamic range requirements. The over-sampling ADC (e.g. the $\Delta\Sigma$ converter), where decimation filter ultimately determines the channel bandwidth, in combination with a switchable analog filter is a promising technique for an SDR, or a reconfigurable, radio.

3.2   Transmitters

The transmitter chain is in many respects the dual of the RX. Also here the trend is towards direct conversion techniques. Figure 2.14 outlines a zero-IF transmitter. Here, the IF signal is a complex single-sideband signal with zero center frequency, generated via a DAC and anti-aliasing filters from look-up tables in the digital baseband. The RF mixer is a quadrature mixer with the LO equal to the RF carrier frequency. Due to imperfections, a co-channel image is generated as well as LO leakage, both degrading the transmitter signal. Nonlinearities and noise in the mixers and IF circuitry together with LO noise

will also widen the transmitted spectrum such that adjacent channels will see some noise. To eliminate this spectrum widening an off-chip RF filter (e.g. SAW) is often inserted in front the power amplifier (PA). Because of stringent noise requirements this TX architecture often has to run at high power levels to suppress circuit-generated noise.
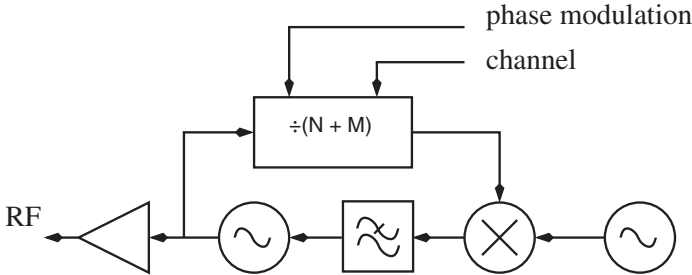
Figure 2.15. Direct phase-modulation transmitter.

By proper choice of waveform tables and DAC dynamic range it is possible to support both nonlinear (e.g. GMSK used in GSM) and linear modulation schemes (e.g. 8PSK used in Edge) in a flexible manner.

For nonlinear modulation schemes, notably GMSK, a very efficient TX architecture has evolved, the direct phase-modulation outlined in figure 2.15. In this case, the transmitter only has to modulate the phase of the RF carrier and not its amplitude (less changing the output power levels off course). Thus, it is possible to use a PLL, where the feedback frequency divider is used to modulate the phase. By using a $\Delta\Sigma$ modulator to "randomly" select the integer divider ratios such that, on average, a fractional divisor results and the truncation noise is pushed outside the loop-filter bandwidth. Because the VCO directly generates the RF carrier very good noise performance can be achieved at a low power consumption, and as only a few high-precision analog components are needed, this architecture is very popular in GSM ASICs. It is sensitive to spuriouses, though, as its reference frequency is only harmonically related to the RF frequency for very few channels. The reference frequency is often 13 or 26 MHz to support sufficient modulation bandwidth.

The direct phase-modulation cannot provide linear modulation as it has no amplitude modulation (AM) possibility. However, by splitting the baseband signal into polar components (ρ–ϕ) rather than Cartesian (I–Q) it is possible to augment the phase-modulator with an AM path, see figure 2.16. When the AM is accomplished via a class-C PA no significant efficiency is gained over the direct-conversion with a linear PA because the amplitude modulation depth has to be absorbed by the AM modulator. However, from a noise perspective the class C saturating PA is better and less filtering is typically needed for this architecture. When the signal bandwidths increase, it becomes increasingly
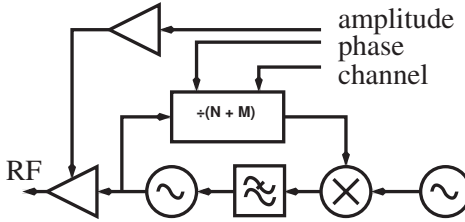
Figure 2.16.   Polar modulation transmitter.

difficult to align the responses of the phase and amplitude paths and, thus, the polar architecture is mostly seen in Edge transmitters. Poor amplitude and phase alignment will widen the transmitted spectrum and this is further aggravated by AM-to-PM and AM-to-AM distortion in the PA. When path alignment and a high-efficiency linear amplitude modulator (e.g. via a fast DC/DC converter) can be accomplished this architecture offers better efficiency than the direct-conversion one.

3.3   Power Amplifiers

A cellular power amplifier has a maximum output power on the order of 25–35 dBm, depending on which standard is targeted. The efficiency is around 50 %, lower for linear modulations and a little higher for GMSK. Because of the modest efficiency and high output power levels, it is difficult to integrate the PA on a transceiver ASIC. For example, if we combine a PA with an LNA on the same die as a WCDMA transceiver we will see more than 130 dB difference between the RF TX carrier power and LNA co-channel noise floor. Furthermore, the active devices in the PA are often built using a non-CMOS technology (e.g. GaAs) making a SOC solution too costly. As integration of more functionality into the PA is desirable, and as matching components, switches, and, bias-control circuits can be put on a SIP, we can have a good mix of technologies in a single part. Because the SIP integrates several of the components around the PA over-all performance optimization is simplified. In particular the TX spectrum during switching or with a mismatched load can be controlled more tightly when the SIP components are co-designed.

4.   Technology Scaling

The driving force behind the technology scaling is the higher gate densities achieved with smaller feature sizes which translates to lower cost. Digital ASICs and memories directly benefit from this and the only major issue with continued scaling is the drain and gate leakage currents, due to lower threshold voltages and thinner gate oxides, respectively. These leakages result in
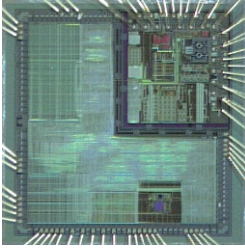
- 0.18 μm CMOS
- high-impedance substrate
- few external components
- 25 % of chip area devoted to RF

Figure 2.17. Single-chip Bluetooth modem.

worse switch on-off current ratios and may ultimately become the dominating contributors to the power consumption of the digital circuitry.

For analog and RF operation the consequences are more involved. For example, high-frequency operation benefit from scaling as the operating frequency is limited by the transit frequency, f_T, see for example [7]. Now, $f_T \propto g_m/C_i \propto I_{dd}/L^2$, that is the maximum operating frequency increases with the bias current and with shrinking dimensions[7]. For example, the switching speed and amplifier gain-bandwidth-product benefit directly from scaling which, in turn, enable better ADCs (higher oversampling ratios and faster active filters). However, the circuit noise floor, v_n^2, is limited by kT/C, see section 2.2, but for bandwidth we have $BW \propto g_m/C$ and $g_m \propto I_{dd}$ which results in $v_n^2 \propto BW/I_{dd}$. Thus, the circuit noise floor does not benefit from scaling. Of course the area of a filter will benefit from higher capacitor densities. The noise figure of an LNA improves with increasing f_T [12], but only to a point after which it is dominated by g_m, and again, the LNA noise figure does not directly benefit from scaling beyond the point when f_T is "high enough".

Device scaling is typically done such that the electrical fields in a device are preserved. The maximum voltage will, thus, track the minimum feature size and as the dynamic range is limited by the battery voltage and the circuit noise floor (i.e. $DR \lesssim V_{batt}^2/v_n^2$) we have to increase the circuit capacitances to preserve the dynamic range when device dimensions and V_{batt} shrink. A net effect is also that the power consumption of analog, dynamic range limited, circuits increases as dimensions are scaled down. So from a traditional analog point-of-view, scaling is a mixed blessing.

The effects of scaling does not differ significantly between CMOS and BiCMOS technologies. Thus, lower supply voltages is a challenge for BiCMOS as well as CMOS designers. In fact, the bipolar base-emitter voltage drop does not shrink with scaling, like the MOS threshold voltage does, making low-voltage BiCMOS design a real challenge.
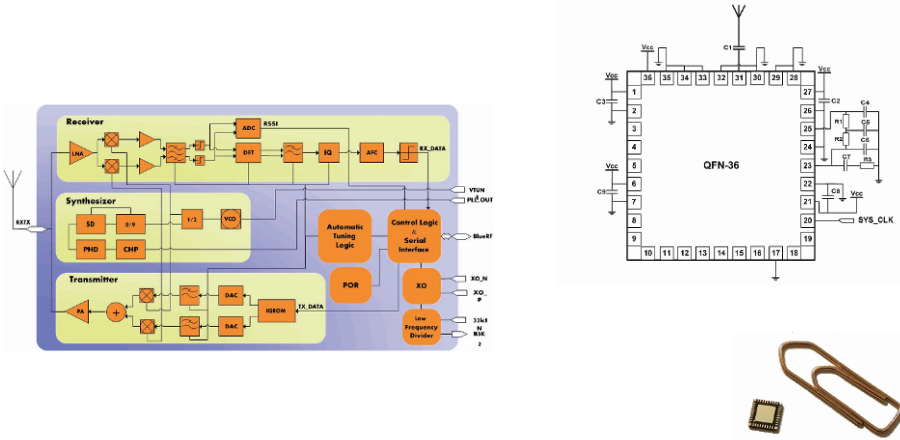
Figure 2.18. Single-chip Bluetooth transceiver.

4.1 CMOS Example

The Bluetooth standard has less stringent RF requirements compared to cellular standards. Much of the relaxation of the standard is due to the use of time-domain duplex, when the transmitter and receiver are operated in non-overlapping time slots which simplifies isolation between these parts of the circuit, higher channel bandwidth which relaxes VCO phase-noise requirements, and, the relaxed receive sensitivity and in-band image suppression which simplifies a SOC implementation. Thus, the very early SOC modems appeared for this standard [4].

In figure 2.17 the chip photograph of a Bluetooth modem [5, 13] is shown. This ASIC uses a standard $0.18\,\mu$m CMOS technology with a high-ohmic substrate and a thick top-level metal. The RF part occupies some 25 % of the chip area which balances the cost between the RF and the baseband. The RF part was further developed into a Bluetooth radio which did not require any external antenna filter nor antenna switch, see figure 2.18.

4.2 BiCMOS Example

Figure 2.19 depicts a triple band GPRS transceiver introduced with the Ericsson T39, circa 2000. This ASIC was implemented both in a bipolar-only and a BiCMOS technology, and, thus, the digital functionality was limited. The RX chain uses a zero-IF architecture with analog I/Q outputs while the transmitter uses direct-phase modulation with the synthesizer frequency divider division ratios calculated by a $\Delta\Sigma$ modulator in the baseband circuit. In-spite the low digital contents, this ASIC achieved a high level of integration with few external components; mainly matching components, PLL loop filter, antenna filters and

- very few external components
- triple band
- on-chip filters
- on-chip VCOs
- bipolar and BiCMOS
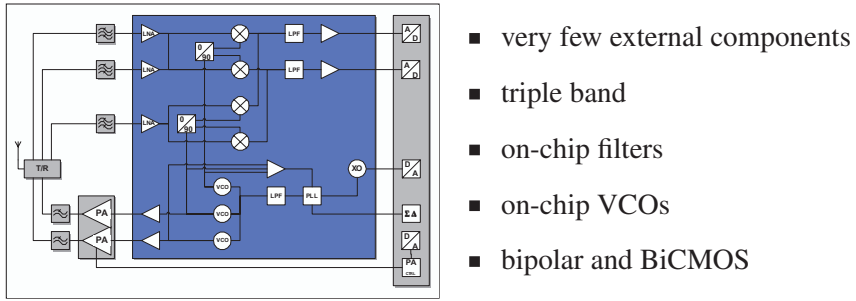
Figure 2.19. Single-chip GPRS transceiver.

switches. In a more recent design only, RX ADC has been added to the receive chain [11] as well as a fourth (850 MHz) band.

5. Handset Implementation Trends

As we have seen in the preceding discussions it is possible to implement a cellular modem with very few components. An increasingly difficult problem is all the internal interfaces in the modem, see figure 2.20. The modem cost is today dominated by the cost of the key components: baseband ASIC, memory, RF transceiver, PA, power management ASIC and the antenna filters and switches. However, for the next few generations the design and specification of these interfaces will be key to a competitive cellular product.

To further reduce the cost of the baseband circuitry, we have to minimize the digital ASIC size and memory requirements (i.e. be flexible in memory configuration such that "just enough" memory can be put in the product) as well as design time. HW reuse (e.g. more use of embedded processors), reconfigurable HW accelerators, and, smaller code size are means to accomplish this. For applications with more moderate volumes, reconfigurable HW, for example using embedded FPGA, can also be envisioned. However, a carefully selected feature set is imperative for obtaining a low over-all cost.

For the RF SOC on-chip inductor area has to be minimized. One inductor corresponds to 10–100 kgates in a deep sub-micron technology. The baseband interface should be digital to minimize pin count and to relieve the digital ASIC from analog I/O blocks. Single-conversion architectures with over-sampled digital signal processing minimize the size of the RF domain and the RF/analog filter needs. Digital trimming can enhance IP_2, compensate for VCO gain variations, reduce DC offsets etc. at a very low area penalty in CMOS. Such trimming willalso improve yield as parametric variations can be reduced.
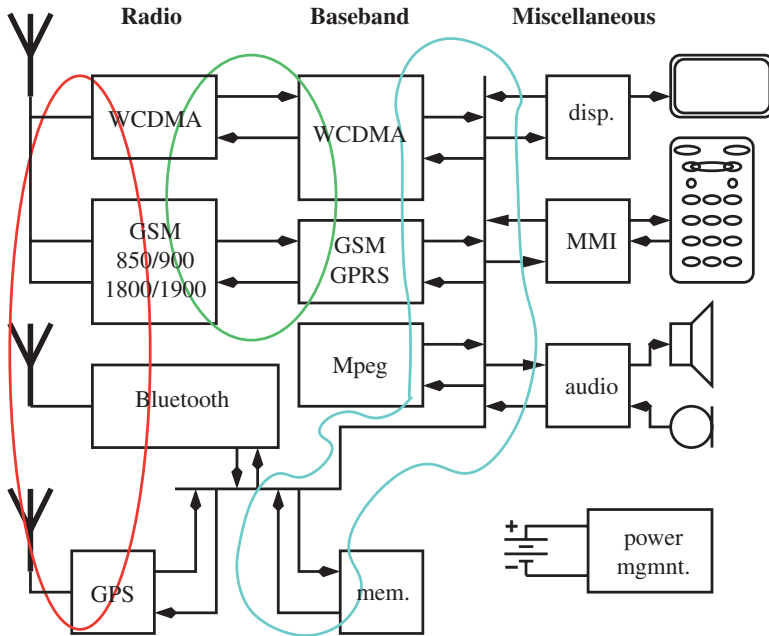
Figure 2.20.   Handset internal interfaces.

Notes

1 Modem is short for modulator-demodulator, and is used for the radio-specific parts of the handset. That is, the signal and protocol processing blocks that convert the digital data and voice bit streams to, and from, radio waves.

2 Temperature compensated crystal oscillator.

3 It is not the capacitor as such that is noisy, it is a reactive component and hence noise free, but when connected between a pair of circuit nodes, there appears to be v_n^2 volts of noise across it.

4 We use GSM as a synonym for both GSM and GPRS.

5 Obviously this will be significantly reduced when the phone is used.

6 The Ericsson GH337, for example, did not use an image filter but a first quadrature mixer which together with the antenna filter provided sufficient image rejection.

7 A quadratic size dependence is only true for long-channel MOS devices. The detailed relation is more complex for deep sub-micron technologies, but a more linear behavior is often observed.

References

[1] Sven Mattisson. Passive integration – an ASIC - SIP comparison. In *RF System in a Package workshop at European Conference on Wireless Technology/European Microwave Conference*, October 2005.

[2] M. Klee, J. van Beek, F. Vanhelmont, F. Roozeboom, P. Lok, F. van Straten, and P. Gamand. System in Package Devices For Mobile Communication and Wireless Data Transfer. In *RF System in a Package workshop at European Conference on Wireless Technology/European Microwave Conference*, October 2005.

[3] K. Muhammad, R. B. Staszewski, and D. Leipold. Digital RF processing: toward low-cost reconfigurable radios. *IEEE Communications Magazine*, 43(8):105–113, August 2005.

[4] F.O. Eynde, J.-J. Schmit, V. Charlier, R. Alexandre, C. Sturman, K. Coffin, B. Mollekens, J. Craninckx, S. Terrijn, A. Monterastelli, S. Beerens, P. Goetschalckx, M. Ingels, D. Joos, S. Guncer, and A. Pontioglu. A fully-integrated single-chip SOC for Bluetooth. In *IEEE International Solid-State Circuits Conference Digest of techical Papers*, February 2001.

[5] P.T.M. van Zeijl, J.-W. Eikenbroek, P.-P. Vervoort, S. Setty, J. Tangenberg, G. Shipton, E. Kooistra, I. Keekstra, and D. Belot. A Bluetooth radio in 0.18 μm CMOS. In *IEEE International Solid-State Circuits Conference Digest of techical Papers*, February 2002.

[6] Asad Abidi. RF CMOS comes of age. *IEEE Journal of Solid State Circuits*, 4:549–561, April 2004.

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

[8] John G. Proakis. *Digital Communications*. McGraw-Hill, New York, third edition, 1995.

[9] G. David Forney. Maximum-Likelihood Sequence Estimation of Digital Sequences in the Presence of Intersymbol Interference. *IEEE Transactions on Information Theory*, pages 363–378, May 1972.

[10] W. Sansen. Distortion in Elementary Transistor Ciruits. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 46:315–325, March 1999.

[11] Y. Le Guillou, O. Gaborieau, P. Gamand, M. Isberg, P. Jakobsson, L. Jonsson, D. Le Déaut, H. Marie, S. Mattisson, L. Monge, T. Olsson, S. Prouet, and T. Tired. Highly Integrated Direct Conversion Receiver for GSM/GPRS/EDGE With On-Chip 84-dB Dynamic Range Continuous-Time $\Sigma\Delta$ ADC. *IEEE Journal of Solid State Circuits*, 40(2):403–411, February 2005.

[12] Osama Shanaa, Ivan Linscott, and Len Tyler. Frequency-Scaleable SiGe Bipolar RF Front-End Design. In *IEEE Journal of Solid State Circuits*, pages 888–895, June 2001.

[13] P.T.M. van Zeijl, J.-W. Eikenbroek, P.-P. Vervoort, S. Setty, J. Tangenberg, G. Shipton, E. Kooistra, I. Keekstra, D. Belot, K. VIsser, E. Bosma, and S. C. Blaakmeer. A Bluetooth radio in 0.18 μm CMOS. *IEEE Journal of Solid State Circuits*, 37(12):1679–1687, December 2002.

Chapter 3

SOFTWARE DEFINED RADIO — VISIONS, CHALLENGES AND SOLUTIONS

Christer Svensson and Stefan Andersson

1.   Introduction

The vision of a software defined radio (SDR) has evolved during recent years [1]. The vision is to have a generic hardware at hand, which can be programmed to perform any kind of radio function and comply with any radio standard. If proposing this idea to a radio engineer he will just find it completely impossible to accomplish. However, technology develops fast, so in our judgement software defined radio will be feasible within 10 years.

So why should we look for a software defined radio? The first proponent was probably the military. They noted that quite many different radio systems are in use today, so if a headquarter needs to keep contact with many operators, it may need many different radios. The trend to have more cooperation between various countries makes the problem even worse. Therefore US military decided to work towards the universal radio, which can talk to any radio system, that is software defined radio. The situation in the civil market is in fact moving in the same direction. The first generation mobile phones had different standards, but the networks were isolated. Then we got the success of the 2G standard, GSM, which in practice became a world standard (even if different bands was used in different countries), and we became used to having a single cell phone working everywhere. Now, new mobile standards emerge, with less success in international standardization, and in the same time we need to use GSM because of its abundance. See also Chapter 1. Furthermore, we have wireless local area networks (WLAN) available in office or at home, and would like to use our cell phone in these networks. The present trend is thus to include 3-4 radios in our cell phone, considerably increasing its price and power consumption. Obviously a software defined radio is strongly needed. We can see a similar

trend for PDA's and laptops; they now communicate with WLAN, but when we are out of reach from a WLAN network we would like to keep connected via the mobile phone network (GPRS, 3G or what may be available). Again, we really need software defined radio. Finally, we have the entertainment sector. Various new broadcasting standards are now in use, DVB-C, DVB-T, DVB-H etc. for digital TV and DAB for digital radio. It would be nice to have a single terminal or set-top box managing all standards, instead of one for each. We also want to have the ability to receive radio and TV in our cell phones and laptops.

In addition to the application needs sketched above, there are many other motivations for software defined radio. Most important is of course cost. SDR facilitates fewer different modules covering market needs, considerably reducing price due to larger volumes. Radio standards are changing, developing, and upgraded continuously. To manage this by just a software upgrade, maybe even in the field, would considerably reduce time and cost, compared to hardware replacement. Of course, development time of new products can be considerably shorter if we can reuse the same hardware and also perform bug fixes in software rather than through a chip respin.

So, if the need is so obvious, why don't we have software defined radios today? The answer is simply that we do not have the technology. Radio design is extremely demanding and the present technique is based on incremental development of traditional radio architectures, with its roots in the beginning of the previous century.

2.   Technical Visions

Most work in SDR has been performed in connection to military demands. Particularly, US government has taken the initiative to develop software platforms for SDR [1]. This research has for example lead to the definition of an open source software communication architecture, Ossie, available from Univ. of Virginia. Ossie is based on CORBA (common object-request-broker architecture) and PoSIX-compliant operating systems. It is however unclear to the authors on what hardware to run this software. In the following we will concentrate on hardware suitable for SDR.

The "ideal" SDR should consist of a programmable digital processor, directly connected to the antenna(s) via AD- and DA-converters. In addition, we most probably need a low-noise amplifier (LNA) and a power amplifier (PA), see Fig. 3.1. The power amplifier could possibly be combined with the DA-converter in the future. The problem with this architecture is of course the very high frequencies at the AD- and DA-converters, asking for sample rates larger than twice the carrier frequencies (Nyquist rate). Traditionally, this problem is solved by using superheterodyne or multiple superheterodyne architectures, see Fig. 3.2. Here the carrier frequency is converted to lower frequencies through one
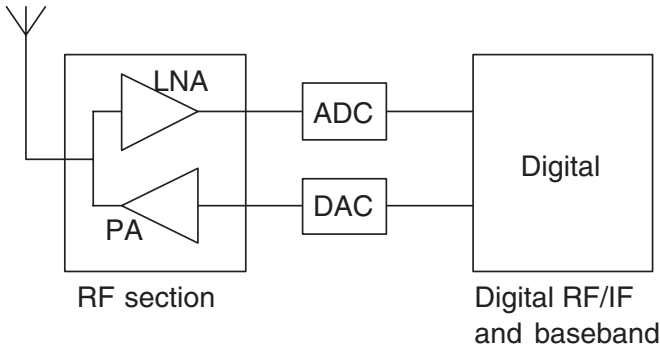
Figure 3.1. "Ideal" software defined radio.
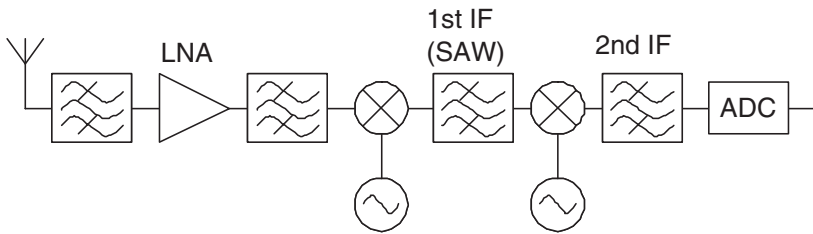
Figure 3.2. Double superheterodyne receiver.

or more mixers, with intermediate filters and amplifiers. The drawback is of course high complexity and many filters, which are not easily programmable.

A more realistic solution is the homodyne or low intermediate frequency superheterodyne, Fig. 3.3. Here we use a single mixer which converts the carrier frequency to baseband or a low intermediate frequency. In such a way we only need a lowpass filter, which is easier to make programmable. The drawback is that the image frequency can not be suppressed by an RF-filter. Instead we use quadrature mixers, which allow image suppression through digital filtering after AD-conversion. The lowpass filter has two roles in this design. First it is an anti-aliasing filter, needed to remove aliasing frequencies occurring because of the limited sampling rate of the AD-converter. Second, it can act as a channel select filter, extracting just one radio channel from the baseband signal. In the second case, it must be programmable if we need to manage channels of different bandwidths (It need to be programmed also in the first case, if we utilize various sampling rates). For a deeper description of conventional RF front-ends, see Chapter 2.
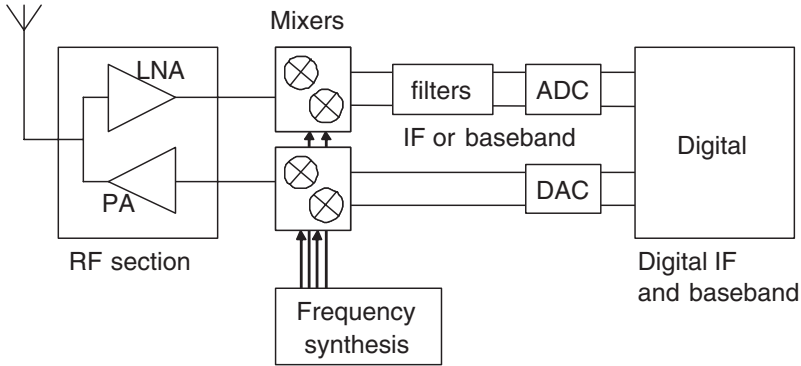
Figure 3.3.   Homodyne transceiver.

For a "true" SDR we want maximum flexibility in carrier frequency and channel bandwidth. We may even wish to be able to manage several different channels in parallel. The best flexibility in the above solutions is obtained by allowing as much signal as possible to reach the AD-converter (receiver side). This is achieved by using only anti-aliasing filtering and by using as high sampling frequency as possible. Still the demands on the anti-aliasing filter may be very high. We will use this "vision" of SDR in the following.

3.   Some Comments on Frequency Planning

In a normal superheterodyne (Fig. 3.4a) we mix the incoming carrier at frequency f_c with a local oscillator of frequency f_{LO}, resulting in an intermediate frequency of frequency f_{IF}, see Fig. 3.4b. Note that because of nonlinearities in the mixers, often nf_{LO} (n natural number) is also active as local oscillator. As shown in Fig. 3.4b, not only signals around the wanted carrier frequency will be converted to f_{IF}, but also some images. To remove these unwanted images, we use an RF-filter. In Fig. 3.4c we show corresponding picture for a homodyne (zero f_{IF}). Here we can not distinguish between the signal and image through an RF-filter, instead we need to use quadrature techniques, as mentioned before.

Quadrature techniques need two local oscillator signals with a phase difference of 90° (sine and cosine signals). An alternative to use two LO signals is to replace the mixer with a sampler, which takes both in-phase (I) and quadrature (Q) samples. I and Q samples are then separated by sorting the stream of samples into one I and one Q stream [2]. This technique is illustrated in Fig. 3.5, where we see how a sampling frequency of $f_s=4f_c/3$ (replacing LO) is used to sample a carrier of frequency f_c, so that every second sample is an I or -I sample and the others are Q or -Q samples. By sorting the samples into two streams
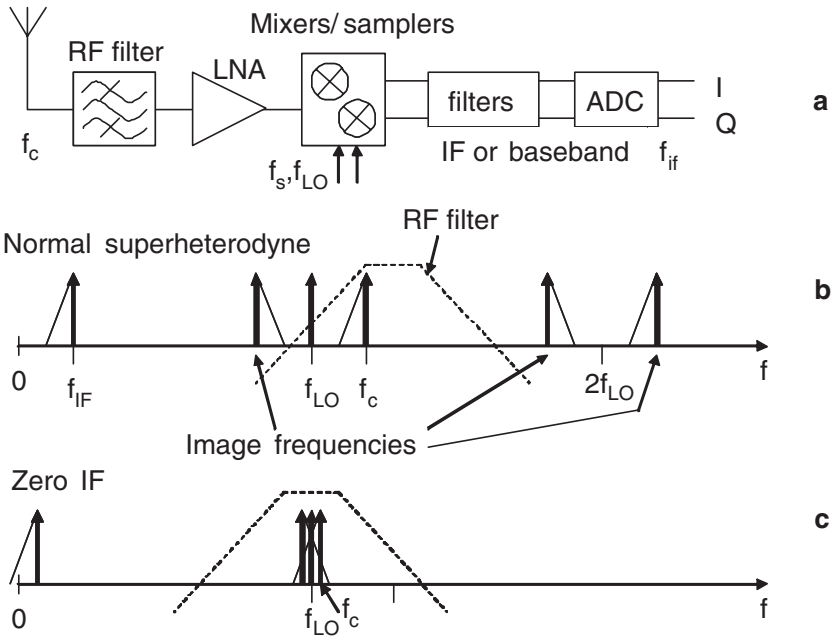
Figure 3.4. Frequency planning for a superheterodyne (b) and a homodyne (c).
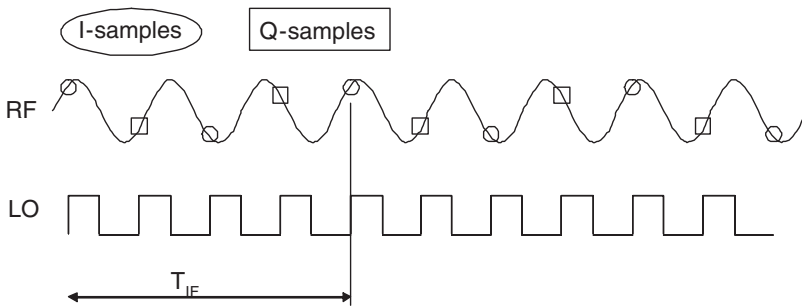
Figure 3.5. I/Q-sampling utilizing $f_s=4f_c/3$.

and multiply every second sample in the two new streams with -1 we obtain one I baseband stream and one Q baseband stream. In general, by choosing f_s according to:

$$f_s = \frac{4f_c}{2m-1}, m \geq 1 \qquad (3.1)$$
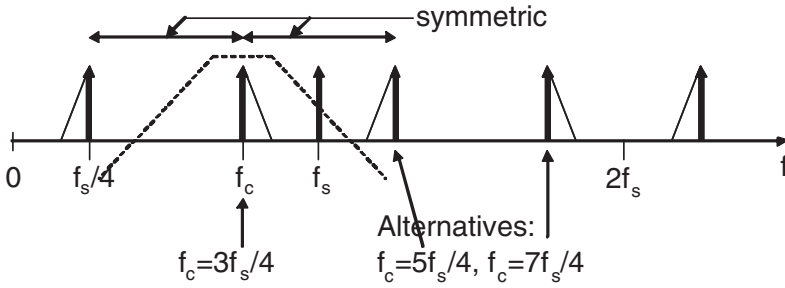
Figure 3.6.   Frequency planning for I/Q sampling with various f_c/f_s ratios.

we get a phase difference between subsequent samples of

$$\Delta\Phi = (2m - 1)\frac{\pi}{2} \tag{3.2}$$

thus making every second sample differ 90° in phase. In Fig. 3.6 we demonstrate the frequency planning for this case, where we demonstrate all possible carrier frequencies which are converted to I/Q baseband for a certain sampling frequency f_s. We note, that there are more images in this case than in the zero IF case, which put more demands on the RF filter (one example of an RF filter shown dashed).

The best flexibility is probably achieved by using a homodyne (zero IF) solution, as it puts a low demand on the RF filter. Instead it asks for two LO signals for I/Q separation. The I/Q sampling solution, on the other hand needs only one LO signal. In order to minimize the demands on the RF filter in this case, we need to use a sampling frequency of $4f_c$ (m=1), which may be a challenge for large f_c.

4.   The Radio Challenge

Radio systems have normally very tough requirements. It is therefore a great challenge to find an alternative architecture from the traditional ones, developed since the beginning of the previous century. Let us here briefly discuss a few of these very tough requirements. We will limit the discussion to the receiver, as the receiver is considered to be the toughest part.

The first large problem in radio systems is the strong disturber (blocker) problem. Assume that we want to receive a signal from a transmitter far away from us; this signal is then very weak. Assume furthermore that we may have another transmitter very close; it will give rise to a very strong signal. Our receiver must then be able to perform an error-free detection of the weak signal in presence of the strong one. We must thus be able to suppress the strong

signal, maybe as much as 10^{10} times (100dB) in order to select the weak one for detection. In addition, if there is more than one strong disturber, any nonlinearity in the receiver front-end will give rise to intermodulation products which may fall into the signal band. Our system must therefore be extremely linear in order not to create such intermodulation. In some cases the strong disturber may be our own transmitter, if both transmitter and receiver are active simultaneously. The analog front-end of the receiver normally manages the strong disturber problem. The principle is to use steep filters at appropriate frequencies to attenuate the strong disturber before it reaches the AD-converter, thus allowing the weak signal to be amplified. The need for steep filters tends to prevent flexibility of the receiver, as it is hard to make steep filters programmable (they are often based on mechanical resonance (SAW filters) or electrical resonators (LC-filters)).

The second large problem is the multipath problem. When we want to receive a signal from a transmitter, it often reaches us through several different paths (direct, and reflected once or several times from various objects). This means that we receive several copies of the signal arriving at different time instances. For short time differences, we may experience interference, which may be constructive or destructive. In the worst case the resulting signal at our antenna may be zero. If we move, the signal strength may vary (often through zero) due to a varying interference. For larger time differences, we may experience a mixing of symbols transmitted at subsequent times. Finally, if we travel with some velocity (like in a car), we will experience a Doppler shift of the transmitted frequencies. All these problems are managed by the digital baseband part of the receiver, and often require very large computing capacity (like the equivalence of 10 Pentium microprocessors for a 3G cell phone) [3]. Today the digital baseband part of the receiver is implemented as custom fixed logic, thus lacking programmability. Programmable digital baseband processing is described in Chapter 5.

Let us return to the analog front-end and roughly estimate the dynamic range needed by an analog to digital converter (ADC) to manage the strong disturber problem (We thus assume that the blocker is not attenuated by filters but reaches the ADC). It is reasonable to assume that we want to be able to detect the weakest signal comparable to the thermal noise of the surrounding:

$$S_n = kT \tag{3.3}$$

Where S_n is the thermal noise spectral density occurring at the receiver, k is Boltzmanns constant and T is absolute temperature of the environment. Taking the low noise amplifier into account, let us instead use the equivalent input noise at the LNA for comparison (that is we include the LNA noise): $S_{ni}=FS_n$, where F is the LNA noise factor. Let us furthermore assume that we should be able to manage a strong disturber 1m away and with 1W output power. We may then

estimate the power of the disturber at the receiver input from:

$$P_B = \frac{A_{antenna}}{4\pi R^2} P_{BTx} \tag{3.4}$$

Where P_B is the blocker power at the receiver, $A_{antenna}$ is the effective antenna area, R is the distance to the transmitter and P_{BTx} is the transmitted power (transmitter assumed to be omnidirectional). We now need to adjust the blocker strength at the ADC, GP_B to the ADC full scale voltage range, V_{FS}, where G is the total front-end gain:

$$GP_B = \frac{V_{FS}^2}{8R_0} \tag{3.5}$$

where R_0 is the impedance level at the ADC input. We can now calculate the quantization noise power of the ADC, P_q [4]:

$$P_q = \frac{1}{R_0} \frac{V_{FS}^2}{12} 2^{-2n} = \frac{2}{3} GP_B 2^{-2n} \tag{3.6}$$

where n is the resolution of the ADC in number of bits. From this we can further calculate the ADC noise spectral density through $S_q=2P_q/f_s$, where f_s is the ADC sampling frequency. Let us now set this quantization noise spectral density equal to the receiver input thermal noise density at the ADC, $S_q=GS_{ni}$. We can then express the requirements on the ADC as:

$$f_s 2^{2n} = \frac{4}{3} \frac{P_B}{FkT} \tag{3.7}$$

The requirements on the ADC can thus simply be expressed in terms of the blocker power at the receiver and the receiver input noise factor! Note that this expression is independent of any choice of radio standard, signal bandwidth, carrier frequency, modulation form etc. In choosing ADC we can trade sampling frequency for bits of resolution as long as we fulfil Eq. 3.7 and the Nyquist criterion for the channel bandwidth. With some reasonable assumptions, we can directly estimate a value of $f_s 2^{2n}$. Choosing F=2 (3dB) and P_B=0.1mW (a 1W transmitter at 1m distance) gives $f_s 2^{2n}=1.6\cdot10^{16}$Hz. In Table 3.1 we show some values of n and f_s which leads to $f_s 2^{2n}=1.6\cdot10^{16}$Hz.

We note that 240GHz is very far from what is available, but 14b, 60MHz is in fact available commercially today. In Fig. 3.7 we show some examples of published AD-converters compared to our requirement [5–7]. So, in principle we should be able to build a software radio based on AD conversion of both blocker and signal today. There are however also other constraints. First, the ADC must fulfil the Nyquist criterion, f_s>2B, where B is the signal bandwidth. For most radios this can be managed (normally B≤20MHz). More problematic is the carrier frequency, which for most radios is f_c≤6GHz. So, either we need

Table 3.1. N, f$_s$ pairs fulfilling our requirement criterion.

n	f_s
8	240GHz
10	15GHz
12	1GHz
14	60MHz
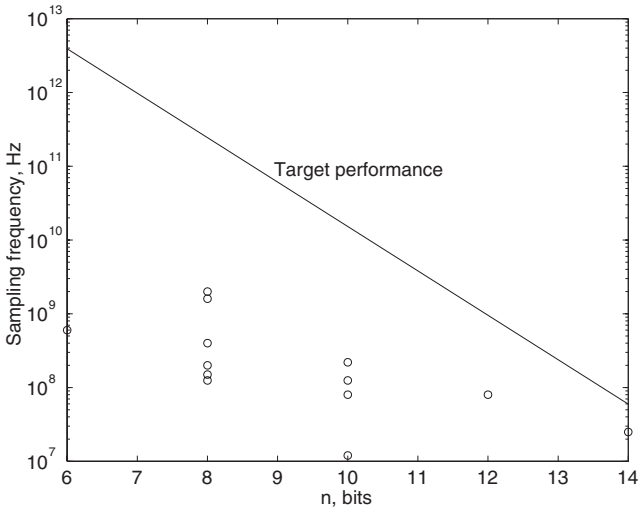
Figure 3.7. Performance of various published ADCs plotted as sampling frequency versus resolution. We also show our target performance $f_s 2^{2n} = 1.6 \cdot 10^{16}$Hz. Data from [5–7].

to fulfil the Nyquist criterion for this frequency, $f_s > 2f_c$, or we need to convert the frequency range from carrier to a low IF or to baseband before AD-conversion. This can be done through subsampling, through a mixer (homodyne) or through sampling plus decimation. In all cases we will create numerous aliasing frequencies, which must be attenuated by filters. This issue is briefly discussed in section 6.

5.   Power Consumption of the Analog to Digital Converter

Coming back to the AD-converter, there is one more problem — the power consumption. In Fig. 3.8 we show the power consumption trend for ADC's, by plotting power consumption versus our requirement measure, $f_s 2^{2n}$. We note that the expected power consumption of an ADC with $f_s 2^{2n} = 1.6 \cdot 10^{16}$ Hz is about 10W. This is obviously far too much for a terminal.
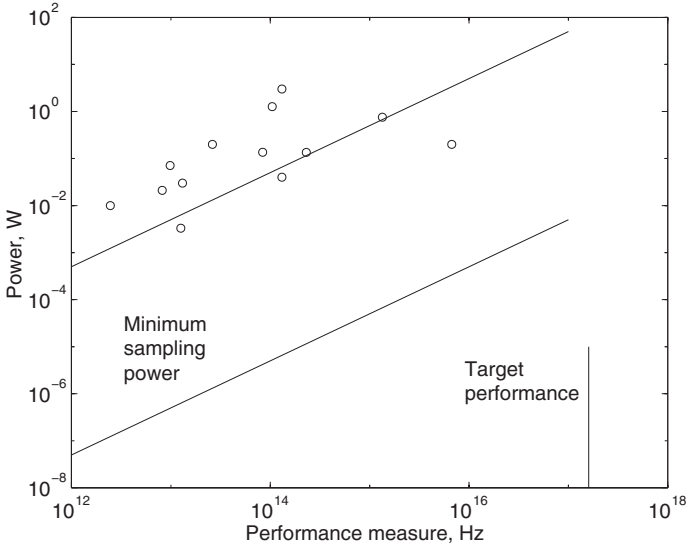
Figure 3.8.   Power consumption versus $f_s 2^{2n}$ for various published ADCs. We also show lines representing $P_S = 12 k T f_s 2^{2n}$ and 10^4 times higher. Data from [5–7].

So what about ADC power consumption? Let us start with a discussion of the minimum ADC power consumption. A simple view is to just consider the power needed for the sampling event [4]. Assuming sampling to take place by a simple switch and a capacitor (Fig. 3.9), we may estimate the thermal noise voltage, V_{ns}, from the switch to $V_{ns}^2 = kT/C_s$ ("kT/C noise"). Making this value equal to the ADC quantization noise (see above) yields:

$$C_s = \frac{12kT}{V_{FS}^2} 2^{2n} \qquad (3.8)$$

The sampling capacitor, C_s, is driven by a maximum current, $I = C_s f_s V_{FS}$, where f_s is the sampling frequency and V_{FS} the full scale voltage, by the previous amplifier. The amplifier must thus deliver a power, P_S, to the capacitor [4]:

$$P_S = I V_{FS} = 12 k T f_s 2^{2n} \qquad (3.9)$$
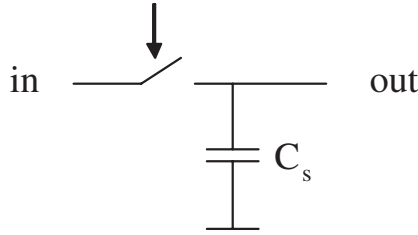
Figure 3.9. Sampling circuit.

where we assumed that also the supply voltage is equal to V_{FS}. P_S is thus the minimum power needed for sampling (and for AD conversion). We note that the minimum power consumption is proportional to the requirements measure, $f_s 2^{2n}$, discussed earlier! For comparison, we plotted P_S and $10^4 P_S$ together with values of actual ADCs in Fig. 3.8. We see that most ADC's have a power consumption that is about four orders of magnitude larger than P_S.

Of course, this very simplified ADC model is not sufficient. Let us therefore discuss the modelling somewhat further. First, is the P_S-modelling above reasonable? If we first check the actual capacitance values given by Eq. 3.8, we can conclude that for a span of n from 8 to 14 bits (and assuming V_{FS}=1V) we get C_s from 3.3fF to 13pF. These values are all reasonable (a minimum transistor gate capacitance in an 180nm process is about 0.9fF and a 13pF MIM capacitor needs about $6000\mu m^2$ area).

In the model above, we assumed that the capacitance is only constrained by thermal noise. What if we need to have capacitor matching (for example we need switched capacitors to control the gain of an ADC stage)? The capacitor must then be large enough so it can be fabricated with enough accuracy. Let us estimate the relative capacitance variance by assuming it equal to the relative variance of transistor β, A_β (both depends similarly on geometry, and A_β is known):

$$\frac{\sigma_C^2}{C^2} = \frac{A_\beta^2}{WL} = \frac{A_\beta^2 \epsilon}{Cd} \tag{3.10}$$

Where W and L are the width and length of the capacitor, and d and ϵ are the dielectric thickness and the dielectric constant respectively. From this we get σ_C/C. The maximum capacitance error, ΔC is then estimated to $\alpha \sigma_C$, where α is a safety margin; let us use α=2 corresponding to a yield of 98% (2% of the capacitors will have an error larger than ΔC). Let us further assume that we need to limit $\Delta C/C$ to one relative LSB, that is 2^{-n}:

$$\frac{\Delta C}{C} = \frac{\alpha \sigma_C}{C} = \alpha A_\beta \sqrt{\frac{\epsilon}{d}} \frac{1}{\sqrt{C}} = 2^{-n} \tag{3.11}$$

From which we can calculate the minimum required value of C:

$$C \geq \frac{\epsilon}{d}(\alpha A_\beta)^2 2^{2n} \tag{3.12}$$

Note that the minimum capacitance is proportional to 2^{2n} as in Eq. 3.8. Using the following parameters for a 0.18μm process: d=50nm, ϵ=3.85ϵ_0 and A_β=0.01μm [8] we find that the minimum capacitance is 5.3 times larger than C_s from Eq. 3.8. The power consumption will thus increase 5.3 times because of this capacitance accuracy requirement. Unfortunately, we have not found any reliable data in literature on how A_β scales.

Furthermore, real AD-converters include much more functionality than just the sampling. If we take a pipelined architecture, for example, it comprises m pipelined stages, each including a comparator and a multiplying sample and hold amplifier [9]. m is somewhat larger than n (redundancy utilized for error correction), but for simplicity we make it n here. Let us start with the comparator. The comparator is in principle a high gain stage. In a clocked comparator, positive feedback is used to make the DC gain infinite. In any case the resolution is limited by the noise level in the beginning of the comparison process. Let us assume a drain noise current spectral density of the first transistor stage of 4kTγg_m (following the standard thermal noise formula for a transistor, where g_m is the transconductance, or the zero drain voltage channel conductance, and $\gamma \approx$1). Further assuming that the bandwidth is given by the load resistance, R_d, and load capacitance, C_d, of the first stage, will give a noise bandwidth of $1/4R_dC_d$. Finally the bandwidth must support the comparator speed, given by the available decision time T_d; we assume R_dC_d=T_d. We may then calculate the drain noise current i_{dn}:

$$i_{dn}^2 = \frac{kT\gamma g_m}{T_d} \tag{3.13}$$

Finally we calculate the equivalent input noise voltage from i_{dn}/g_m and equalize this voltage to the quantization noise of the AD-converter, R_0P_q (Eq. 3.6) and obtain a minimum g_m to fulfil the noise criterion:

$$g_m \geq \frac{12kT\gamma}{T_dV_{FS}^2}2^{2n} \tag{3.14}$$

In order to reach this required g_m we need a drain bias current of:

$$I_D = g_mV_{eff} \tag{3.15}$$

Where V_{eff}=(V_G-V_T)/2 for an MOS transistor, where V_G is the gate bias voltage and V_T is the threshold voltage. From this we estimate the minimum comparator power from I_DV_{FS} (assuming again a supply voltage of V_{FS}):

$$P_C = \frac{12kT\gamma V_{eff}}{T_dV_{FS}}2^{2n} \tag{3.16}$$

We note that this formula is very similar to the minimum sampling power formula (Eq. 3.9), so P_C is of the same order of magnitude as P_S.

For the multiplying sample and hold amplifier we assume a single stage OTA in a switched capacitor context. Our accuracy assumption is that the output must settle to 2^{-n} in settling time T_s. Also, the circuit must have low enough noise. The capacitors sizes, C_L, must fulfil the kT/C noise criterion (Eq. 3.8). This capacitor is also the load for the OTA. We then estimate the settling time constant of the OTA to g_m/C_L, where g_m is the transconductance of the OTA (and of the OTA input transistor). Equalizing the settling error to 2^{-n} gives:

$$e^{-\frac{g_m T_s}{C_L}} = 2^{-n} \tag{3.17}$$

From which we can calculate a minimum g_m as above, and a minimum supply current:

$$I = \frac{nC_L V_{eff} ln2}{T_s} = \frac{12kT}{T_s V_{FS}} 2^{2n} \frac{V_{eff}}{V_{FS}} nln2 \tag{3.18}$$

And finally we calculate the minimum power:

$$P_{OTA} = IV_{FS} = \frac{12kT}{T_s} 2^{2n} \frac{V_{eff}}{V_{FS}} nln2 \tag{3.19}$$

Again we see a large similarity to P_S (Eq. 3.8), in this case P_{OTA} is roughly n times larger than P_S. Note that taking capacitance accuracy into account (Eq. 3.12) will increase the power consumption by the same amount as above ($\sim$5.3x for a 0.18μm process). In the above estimation we did not consider slewing time. We expect settling time to dominate over slewing time for large n (say $n\geq 8$).

Let us finally combine these results into an estimation of the power consumption of an n-stage pipelined AD-converter. Such a converter thus includes n comparators and n operational amplifiers and will have the minimum power consumption of:

$$P_{pipe} = n(P_C + P_{OTA}) = IV_{FS} = 24nkT f_s 2^{2n} \frac{V_{eff}}{V_{FS}} (nln2 + \gamma) \tag{3.20}$$

Where we used $T_d = T_s = 1/2f_s$. Note that again the power consumption is proportional to our requirements measure, $f_s 2^{2n}$! Taking n=10, f_s=40MS/s, V_{eff}=0.05V and γ=1 as an example, we arrive to a minimum power of P_{pipe}=17μW (compared to P_S=2μW). Taking capacitance matching into account the power figure will increase to 79μW.

As a result from the analysis above we draw the conclusion that the minimum power consumption of a 10 bit ADC is about 10 times larger than P_S or, if we consider capacitor accuracy, about 40P_S.

Regarding accuracy, we believe that accuracy of the AD-converter building blocks can be completely replaced by digital error correction in the future (digital error correction is utilized to a relatively high degree already today [9]). Thus we do not need to consider the capacitor accuracy discussed above. Also, we do not need to consider offset voltages. We may even remove the requirement of settling used above; however then we will instead have a slewing problem, so we will probably come up with about the same minimum power.

In conclusion, we expect that it will be possible to reduce AD-converter power consumption to a level of the order of $10P_S$, where P_S is given by Eq. 3.9. We further note that the estimated minimum power do not include any transistor process parameters, thus we have not considered effects of transistor speed limitations. We can therefore expect that our estimations only are valid for limited sampling frequencies (compared to transistor f_T). Finally, estimating the minimum power needed to fulfil our SDR requirement of f_s2^{2n}=1.6·10^{16}Hz we arrive at an AD-converter power consumption of about 10mW, which is a reasonable value.

Some recently announced AD-converters supports this conclusion. TI recently announced an AD-converter with n=14 and f_s=190MS/s with a power consumption of 1.1W (ADS5596). This gives a power consumption of $430P_S$, only 40x the minimum estimation above. Another example, Linear Devices recently announced a device with n=16 and f_s=130MS/s with a power consumption of 1.25W (LTC2208). This corresponds to $40P_S$, which is just 4x our minimum estimation!

6.   Other Key Components

Above, we concentrated our discussion on AD-converters. However, in most cases it is not realistic to have the ADC sample the RF signal directly at Nyquist rate (>2f_c). We therefore need to reduce the frequency to be converted or we need to utilize subsampling. In any case, mixing, subsampling, or sampling followed by decimation will create image and alias frequencies. We therefore need to perform filtering in the RF chain. If we want to cover a very large range of RF frequencies, we thus need to have widely tunable RF filters. One way to solve this is to have a bank of filters, which can be switched into the signal path. Another solution is to utilize widely tunable filters, for example based on g_m/C type recursive filters [10]. None of these techniques are mature today. The RF filter normally removes images caused by the mixing process. In order to remove alias frequencies caused by the low sampling frequency of the AD-converter, we need a steep low-pass filter in front of the ADC. In a homodyne, such filters are normally realized by active g_m/C filters. In sampling receivers switched capacitor filters are more convenient, as the signal already is discrete time. See the second example in section 8. Particularly elegant is to design a filter with notches at the unwanted aliasing frequencies [11].

Table 3.2.   Requirements of ISM receiver.

Parameter	Specification
Frequency band	159±4MHz
Channel bandwidth	25kHz
Receiver sensitivity	-104dBm
Blocker power	-15dBm
Two-tone test power	-27dBm

Furthermore, normally the input signal is very weak, so we need a low-noise amplifier (LNA) in front of the mixer/sampler/ADC. The very high demand on dynamic range puts tough requirements on the linearity of the LNA, requirements which often are quite hard to fulfil (see section 4). Also, if a strong blocker will reach the LNA input, there is not much room for gain in the RF chain, so we need to have low noise not only at the LNA input but also in mixer/sampler and in anti-aliases filters.

7.   Example of a 160MHz Carrier SDR Front-End

A particular application needs three different receivers in the AIS band (159±8MHz) to be operated in parallel. We then tried a solution where the three superheterodynes in the conventional solution are replaced by a single RF front-end [12]. The requirements on the receiver are shortly described in Table 3.2. Our proposed solution is demonstrated in Fig. 3.10. The receiver chain is thus composed of a wide-band LNA, an RF-filter, an automatic gain controlled
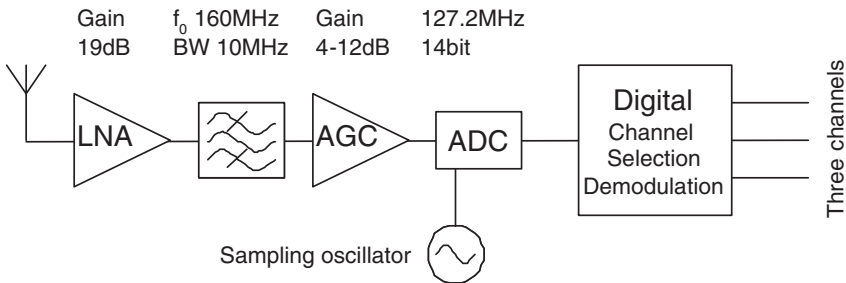
Figure 3.10.   SDR implementation of a three channel radio for AIS band.

amplifier (AGC), and a 14 bit ADC. We utilize I/Q subsampling at f_s=4f_c/5 (m=3 in Eq. 3.1), where the sampling is performed by the AD-converter itself. I/Q separation thus takes place in the digital domain after AD-conversion. Also, the three desired channels are separated in the digital domain utilizing digital mixers and filters.

Let us discuss how the requirements are met in this receiver. First, we chose a sampling frequency of 127MHz (corresponding to a nominal carrier frequency of 158.75MHz). As the channel bandwidth is 25kHz, this sampling frequency corresponds to an oversampling ratio (OSR) of 2540, contributing an oversampling gain of 34dB. The worst case dynamic range occurs for a signal of -104dBm and a blocker at -15dBm and with an SNR requirement of 12dB (FSK modulation) and becomes 101dB. The oversampling gain of 34dB leaves 67dB dynamic range requirement on the AD-converter. This corresponds to 11 bit (and $f_s 2^{2n}$=5·10^{14}Hz). However, a 12 bit ADC do not normally reach 67dB dynamic range at its maximum sampling frequency, so we have chosen a 14 bit ADC. Furthermore, we also need a linearity corresponding to an intermodulation suppression of -27+104+12=89dB, which motivates the choice of a 14 bit ADC. The closest images in this case occurs at 3f_s/4 and 7f_s/4, that is 95.25MHz and 222.25MHz respectively. This asks for quite a steep RF-filter, so we used an off-the-shelf filter centered around 160MHz and with a 3dB bandwidth of 10MHz (to include the whole ISM band) and a 60dB bandwidth of 32MHz. Finally, we choose the RF chain gain so that the blocker voltage corresponds to the ADC full scale voltage at ADC input.

Measurements were performed on an experimental front-end. Output data from the ADC was collected into a computer and the following signal processing performed off-line in MATLAB. In Fig. 3.11 we show an example of an unfiltered digital signal, with a blocker of -15dBm (at frequency offset 290kHz) and a signal of -71dBm (at frequency offset zero). We still have a signal-to-noise ratio of about 14dB around the signal (the blocker can be suppressed by a digital filter). The minimum signal handled by the receiver was -95dBm and the maximum -7dBm, leading to an observed dynamic range of 88dB. We thus did not reach full performance in this experiment, but the experiment clearly indicates the possibilities.

8.   Example of a 2.4GHz Carrier Front-End

The second example was designed for a 2.4GHz carrier for WLAN, but can be run at any carrier frequency up to 2.4GHz [13]. The goal was to have a bandwidth of 20MHz (at 2.4GHz carrier) and a maximum ADC sampling rate of 100MS/s. The proposed solution is shown in Fig. 3.12. We will here only discuss the RF sampling downconversion filter, which was realized as an experimental chip. Here we again utilize I/Q subsampling, at f_s=4f_c/9 (1072MS/s at 2.4GHz carrier). The stream of samples are sorted into one I and one Q stream.
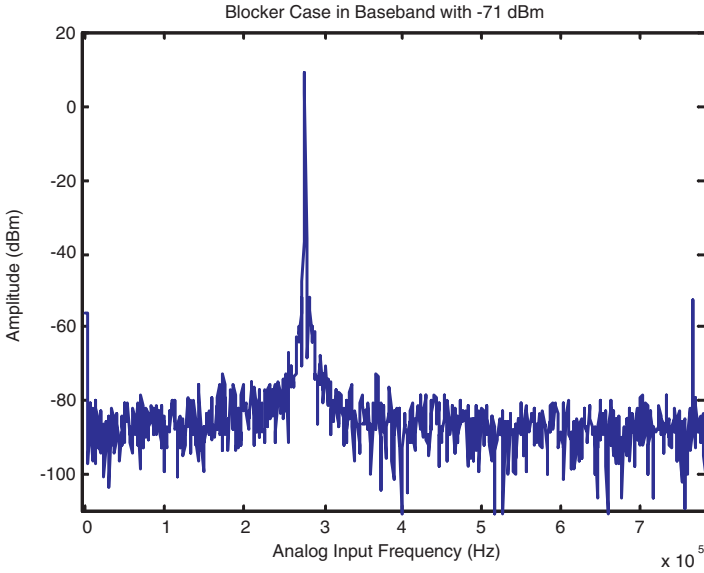
Figure 3.11. Unfiltered digital signal including a signal at zero and a large blocker at 290kHz.

Each of these streams are then decimated and filtered in a switched capacitor decimator/filter. The principle is to take the weighted average of 12 samples to perform an FIR filter (see Fig. 3.12). From a FIR filter perspective, each filter have a length of 23, but with every second coefficient zero (corresponding to
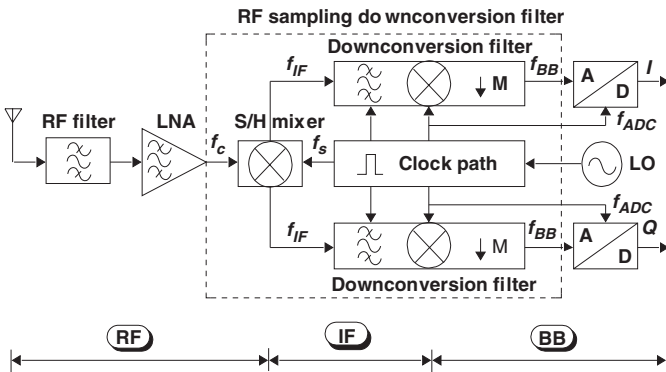
Figure 3.12. Block diagram of an RF front-end based on a sampling downconversion filter.
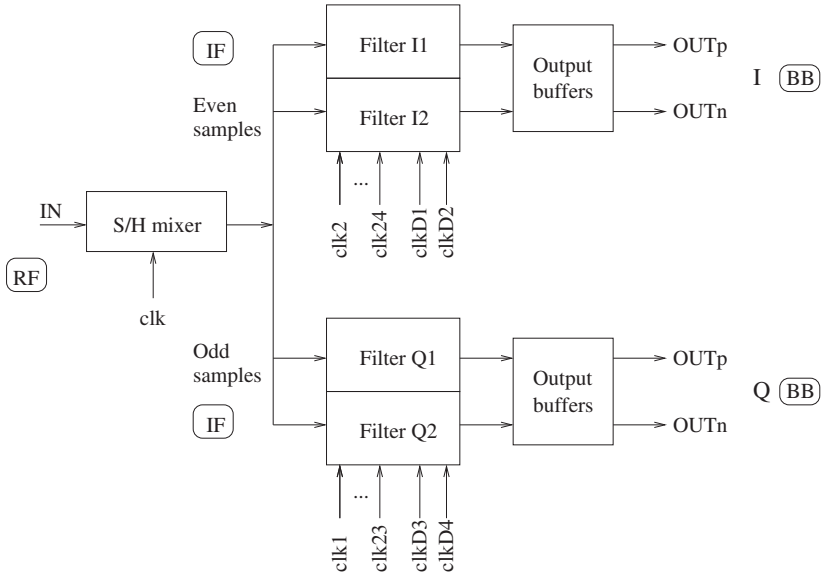
Figure 3.13.   Implementation of the filter in the sampling downconversion unit.

Table 3.3.   Non-zero coefficient multiplication and summation (S) sequence for the filter implementation shown in Fig. 3.13.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Filter I1	S	a_2		a_4		a_6		a_8		a_{10}		a_{12}		a_{14}		a_{16}		a_{18}		a_{20}		a_{22}		S
Filter I2	a_{12}	a_{14}		a_{16}		a_{18}		a_{20}		a_{22}		S	a_2		a_4		a_6		a_8		a_{10}		a_{12}	
Filter Q1	S	a_2		a_4		a_6		a_8		a_{10}		a_{12}		a_{14}		a_{16}		a_{18}		a_{20}		a_{22}		
Filter Q2	a_{12}	a_{14}		a_{16}		a_{18}		a_{20}		a_{22}		S	a_2		a_4		a_6		a_8		a_{10}			

the samples in the other data stream). The decimation factor is 12, so that the output sampling frequency is 1072/12=89MS/s. As the filter length is about twice the decimation factor, we need two interleaved filters for each I and Q data stream, see Fig. 3.13. Table 3.3 shows the coefficient multiplication (with a_i) and summation (S) sequence as a function of input sample for each of the four filters in Fig. 3.13. The two outputs from each filter is then multiplexed together in the output buffers. Furthermore, in this implementation positive and negative filter coefficients are separated and the final subtraction is performed
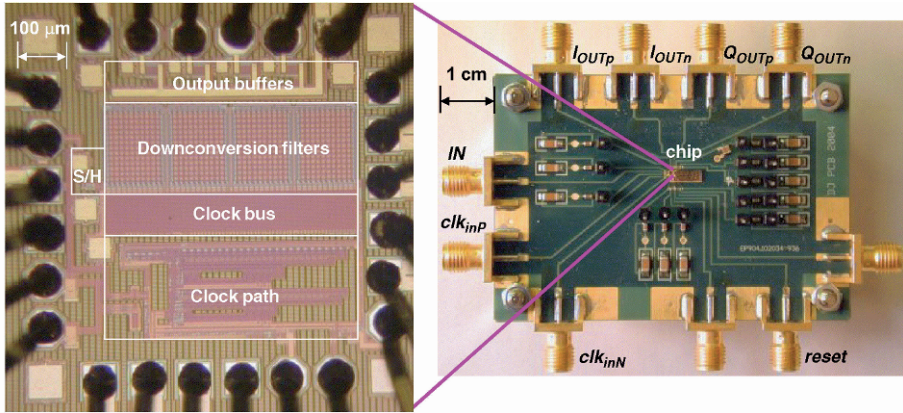
Figure 3.14. Chip photo of the sampling downconversion filter (a) and test board (b).

at the AD-converter, utilizing its differential input (this to avoid subtraction to be performed in the switched capacitor circuit).

The RF sampling downconversion filter was implemented in a $0.18\mu m$ CMOS process, see chip photo in Fig. 3.14a. The experimental chip was bonded directly to a printed circuit board for testing, Fig. 3.14b. Measurements verified the functionality of the chip. In Fig. 3.15 we show the filter response, indicating the channel bandwidth, BW_{ch}. The first aliased band is also shown, with an aliasing rejection of about 17dB. In Table 3.4, we list all relevant properties of this chip. We also ran the chip with real modulated data and measured the constellation diagram for a 64QAM modulated carrier, see Fig. 3.16. We note that the I/Q accuracy is sufficient to resolve this high modulation index.

In conclusion, sampling mixers combined with discrete time decimation and filtering is a very promising alternative to the traditional homodyne. Such circuits are easily made flexible (running at various frequencies) and they are also very robust. In addition to the example above, TI has published several similar circuits [14, 15].

9.   Conclusion

At first sight, software defined radio (SDR) appears to be far away in the future. However, a more careful analysis indicates great opportunities. It is the authors' opinion that SDR will be used in commercial products before 2015. We can distinguish three challenges, the RF front-end, the AD-converter and the digital signal processing. Regarding the RF front-end, both the homodyne and the sampling solutions are very promising, with several working demonstrators. Also, new principles for tunable RF-filters/LNAs have been demonstrated. Regarding the AD-converter, recent progress strongly indicates that converters
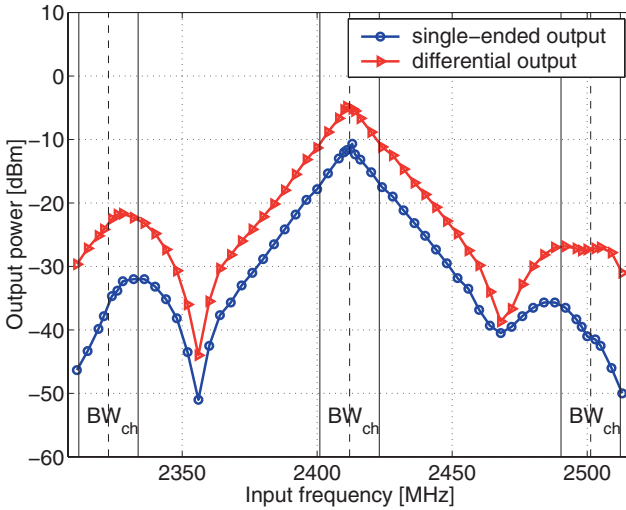
Figure 3.15. Measured filter response, single-ended output and differential output. BW$_{ch}$ is the anticipated channel width.

Table 3.4. Measured data of the sampling downconversion filter for two different sampling frequencies.

Parameter	Specification		
Nominal center frequency		2412MHz	
Input sampling rate	1072MS/s		567.5MS/s
Output sampling rate	89.3MS/s		47.2MS/s
Gain	-1dB		1dB
Alias band rejection	>17dB		>24dB
Image rejection	59dB		29dB
Noise PSD	-131dBm/Hz		-130dBm/Hz
Jitter	0.64ps		0.54ps
Analog input bandwidth		4.55GHz	
Supply voltage		1.8V	
Power consumption	87mW		70mW
Core area		0.36mm^2	
Technology		0.18μm CMOS	
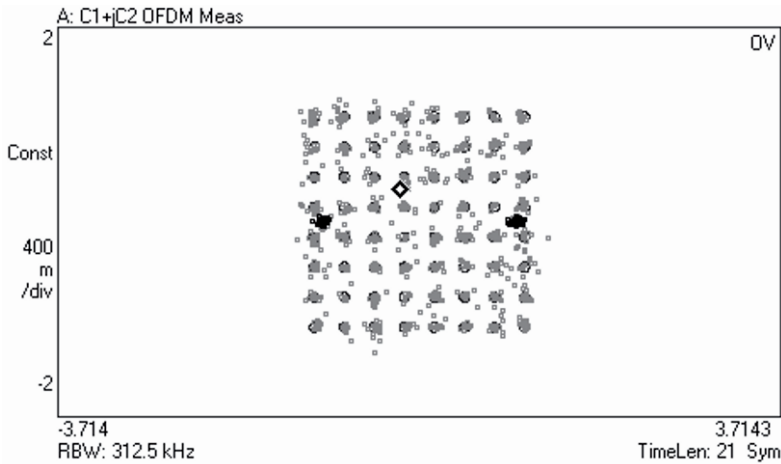
Figure 3.16. Measured constellation diagram when receiving a 64QAM modulated signal.

with sufficient dynamic range for converting both signal and blocker can be made with low enough power consumption. Finally, recent progress in digital baseband signal processors shows that the digital processing problem for SDR is close to being solved.

References

[1] D. Marsh, Software-defined radio tunes in, *EDN*, March 2005, p. 52.

[2] R. G. Vaughan, N. L. Scott and D. R. White, The theory of bandpass sampling, *IEEE Transactions on Signal Processing*, vol. 39, pp. 1973-1984, Sept. 1991.

[3] C. H. van Berkel et. al., Vector Processing as an enabler for Software-Defined Radio in Handsets from 3G+WLAN Onwards, *The 2004 Software Defined Radio Technical Conference*, Scottsdale, AZ, 16-18 Nov. 2004.

[4] P. B. Kenington and L. Astier, Power Consumption of A/D Converters for Software Radio Applications, *IEEE Trans. On Vehicular Technology*, vol. 49, pp. 643-650, March 2000.

[5] Papers 4.2, 14.1, 14.2, 14.3, 14.4, 14.5, 14.7, and 25.5, *in Proc. IEEE International Solid-State Circuits Conference*, Feb. 2004.

[6] Papers 15.2, 15.4, and 15.5, *in Proc. IEEE International Solid-State Circuits Conference*, Feb. 2005.

[7] F. Vessal and C. A. T. Salama, An 8b 2-Gsamples/s Folding-Interpolating Analog-to-Digital Converter in SiGe Technology, *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 238-241, Jan. 2004.

[8] P. R. Kinget, Device Mismatch and Tradeoffs in the Design of Analog Circuits, *IEEE Journal of Solid-State Circuits*, vol. 40, pp. 1212-1224, June 2005.

[9] S. H. Lewis, Optimizing the Stage Resolution in Pipelined, Multistage, Analog-to-Digital Converters for Video-Rate Applications, *IEEE Transactions on Circuits and Systems II*, vol. 39, pp. 516-523, Aug. 1992.

[10] S. Andersson and C. Svensson, A 750MHz to 3GHz Tunable Narrowband Low-Noise Amplifier, *Proc. of the NORCHIP 2005 Conference*, pp. 8-11, Nov. 2005 .

[11] S. Andersson, J Konopacki, J. Dabrowski and C. Svensson, SC-filter for RF Sampling and Downconversion with Wideband Image Rejection, *Accepted for publication in Analog Integrated Circuits and Signal Processing*.

[12] S. López, S. Otero and C. Svensson, Direct Sampling Receiver Front-end for the VHF band, *Proc. of RadioVetenskap och Kommunikation*, pp. 253-256, June 2005.

[13] D. Jakonis, et. al., A 2.4-GHz RF Sampling Receiver Front-End in 0.18μm CMOS, *IEEE Journal Solid-State Circuits*, vol. 40, pp. 1265-1277, June 2005.

[14] K. Muhammad, R. B. Staszewski and D. Leipold, Digital RF Processing: Towards Low-Cost Reconfigurable Radios, *IEEE Communications Magazine*, p. 105-113, Aug. 2005.

[15] K. Muhammad, et. al., A Discrete-Time Bluetooth Receiver in a 0.13μm Digital CMOS Process, *Proc. IEEE International Solid-State Circuits Conference*, pp. 268-269, Feb. 2004.

PART II

DIGITAL SOC DESIGN

Chapter 4

TRENDS IN SOC ARCHITECTURES

Ahmed Hemani and Peter Klapproth

1. Introduction

Trends in SOC Architectures are shaped by the demands of applications on performance, cost and power consumption and the developments in VLSI, Design and Battery technologies.

The application space for SOCs is wide and varied. The discussion in this chapter primarily focuses on applications in the wireless multi-media domain. This narrowed application space can be divided into two broad categories. One related to communications systems and the other related to multi-media applications.

In communication systems, linear increase in bandwidth requires exponential increase in computational power. Ample evidence of this phenomenon was presented by Jan Rabaey [1]. One graph from this presentation, due to Ravi Subramanian, Morphics, is shown in Figure 4.1, which depicts that the algorithmic complexity increases three orders of magnitude between successive generations of the wireless standards. This increase in complexity comes from attempting to reach the Shannon limit in dealing with non-ideal channels. VLSI technology provides the computational power for implementing the communication systems. If a hypothetical DSP processor that had the computational power to implement the communication system algorithm(s) in 1980 corresponding to 1G in Figure 4.1, then the gap between improvement in raw computational power of the same DSP processor as VLSI technology progresses following Moore's law and the increase in *average algorithmic complexity* of communication systems at subsequent corresponding points in time is widening. To
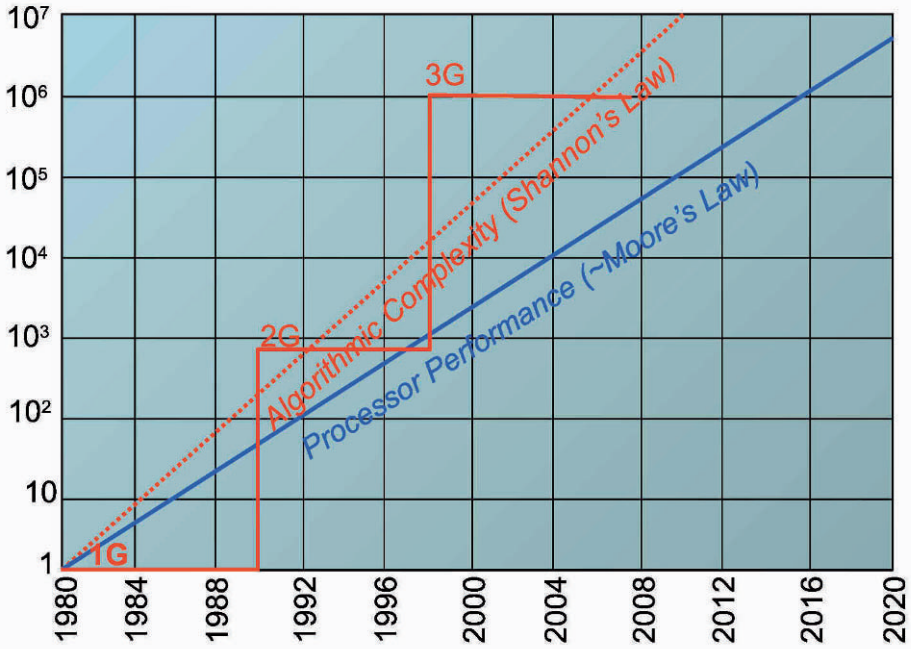
Figure 4.1.   Shannon beats Moore. Source Rabaey [1].

bridge this widening gap, VLSI systems *evolve architecturally* to meet the computational demands of the increased algorithmic complexity. For this reason, this gap is called *the architectural efficacy gap*.

A further nuance in this description is that while the raw computational power increases at a more fine grained granularity, every 18 months or so, following Moore's law, the algorithmic complexity of communication systems increases in bursts. Communication standards, once defined, tend to remain the same along with their algorithmic complexity for a long time to recover the enormous investment made in the infrastructure for a particular standard. This means that while our hypotetical DSP processor from the 1G era is far short of computational power, in its original architecture, at around 2000 to meet the algorithmic complexity of 3G systems, the same architecture would grow in raw computational power around 2015 to match the computational complexity of the 3G systems.

The gap discussed above and shown in Figure 4.1 is further aggravated by the fact that wireless devices today sport several radios like WLAN for high bandwidth low cost internet access, Bluetooth for handsfree headsets, GPS for navigation, etc. Multi-media applications like mega pixel camera, audio and
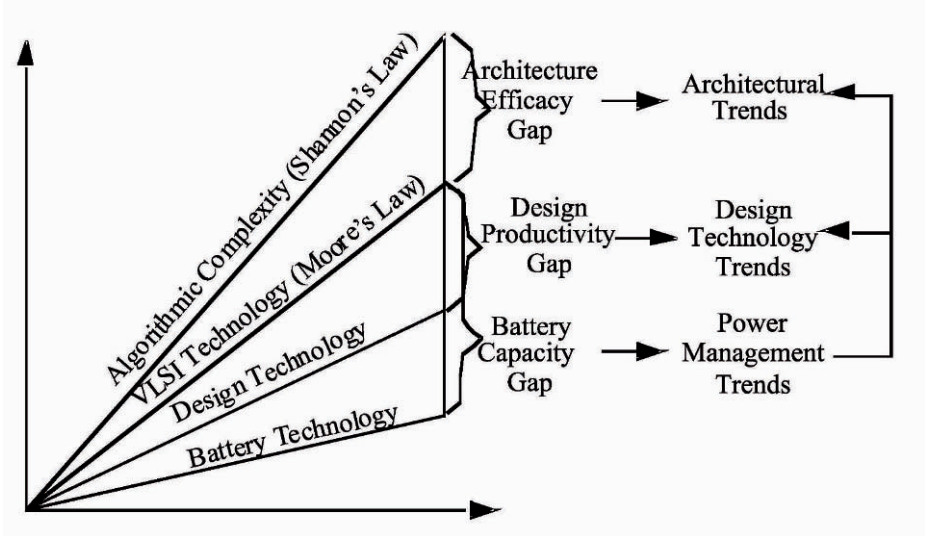
Figure 4.2. Growth in Algorithmic Complexity vs Progress in VLSI, Design and Battery Technologies.

video streaming (see [2]) are all adding up to a performance requirement that clearly outpaces the improvements in raw computational power provided by progress in VLSI technology.

Another critical technology that is not keeping pace with the demands is battery technology. Improvement in battery technology are growing at an even slower pace compared to VLSI technology. To bridge the increasing gap between power requirement and availability for battery powered systems and the limits on dissipation, VLSI Systems have once again responded by evolving architecturally to bridge the *power capacity gap*.

A third gap, the *design productivity gap*, results from slower than required development of the design technology compared to the development of the VLSI technology. Like algorithmic complexity, design technology improves in bursts, and according to the author [3] on average matches Moore's law.

These gaps have a strong influence on the Architectural and Design Technology trends of SOC. Design Technology trends have been discussed in detail in [3], in this chapter architectural trends are discussed.

2.   VLSI Design Space

SOC design has three obvious goals: higher performance, low power consumption and low cost. These goals are partly achieved by technology scaling. As technology scales, device area shrinks quadratically, device propagation delay under the constant field assumption reduces linearly and the interconnect delay scales negatively. Supply voltage reduces less than linearly. This raw and natural improvement with technology scaling, as argued above, is not enough to keep pace with demands from the application space. This is especially true for performance and power consumption. Performance improves inversely with reduction in combined propagation delay in device and interconnect. As interconnect delay begins to dominate the overall delay, performance improvement due to technology scaling is progressing at much slower pace. Dynamic power consumption scales down quadratically with supply voltage but because of supply voltage standard and performance requirement, industry has lowered supply voltage at a slower pace than technology scaling. Moreover, lowering supply voltage, while reducing the dynamic power consumption, indirectly contributes to increase in static power consumption. To maintain performance, reducing supply voltage also requires reducing threshold voltage which increases the leakage current. Leakage current went up from 20 $\rho A/\mu m$ in TSMC $.18\mu m$ with a threshold voltage of 0.42V to 13000 $\rho A/\mu m$ in TSMC $.13\mu m$ with a threshold voltage of 0.25V according to [4].

SOC architectures have evolved to bridge the gap between improvements in performance and power consumption that can be achieved by technology scaling and what is required. This evolution happens in the three dimensional design space of Area, Delay and Power. Flynn et. al. in [5] have presented bounds on this space by showing pairwise relationship between the three parameters as shown in Figure 4.3. The area performance relationship is an exponential one. The exponent n is between 1 and 2 as presented in [5]. In other words, to double the performance the area needs to be doubled in the best case and quadrupled in the worst case. Increasing area adversely affects power consumption because the static power consumption is directly proportional to the total gate width of the design. The power performance is cubically related and requires expending eight times more power to double the performance. And as technology scales, these relationships scale parametrically as shown in Figure 4.3. In the next few sections we discuss the impact of architectural evolution on the components of SOC design, i.e., embedded processors, memories, interconnect and power management schemes.

2.1   Trends in Embedded Processor Architecture

Embedded processors can be divided into two broad categories (see [6]). The first category is often referred to as micro-controllers that typically deliver
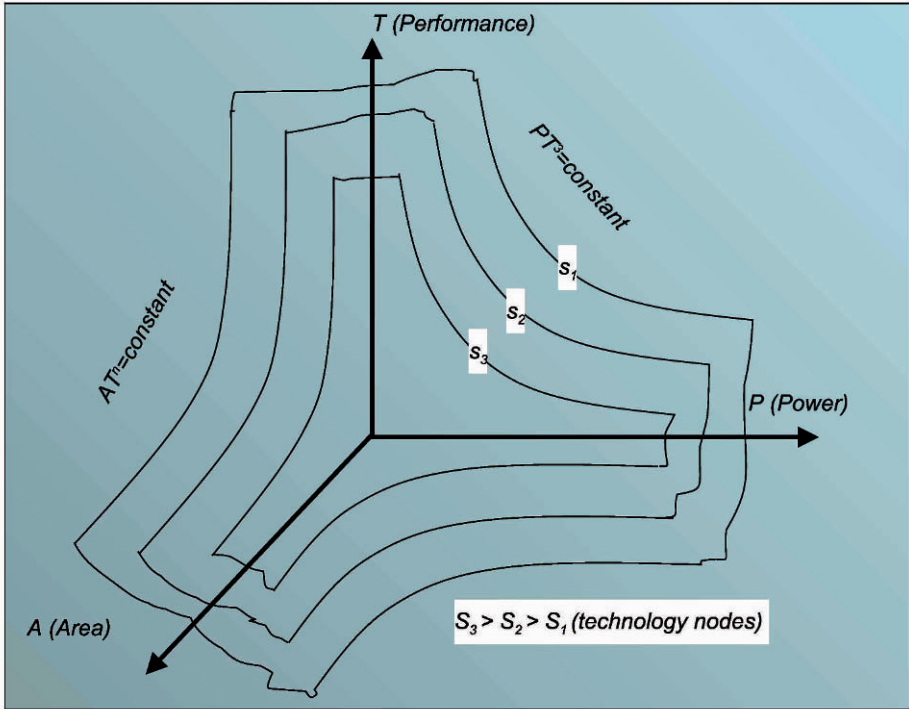
Figure 4.3. VLSI Design Space (Source Flynn et. al. [5]).

20 to 150 MIPS and integrate RAM and ROM but do not contain caches. This category targets control functionality in VCRs, Washing Machines etc and peripherals like hard disks. This first category can by expanded to include low end cache less DSPs used for control and light weight signal processing algorithms. The second category delivers 100-500 MIPs, integrates caches and can have advanced DSP and Multi-media instructions. This category finds application in advanced mobile-phones, cameras etc. Embedded processors have always been the volume leader in the processor market and the trend continues. According to the SIA roadmap, embedded processors that are integrated in SOCs are increasing at an annual rate of 20% whereas the compute intensive high-end desktop variants are increasing at 8%. Embedded processors also have a much tougher power budget compared to their desktop cousins. The SIA power roadmap predicts a less than fifty percent increase in power consumption of portable embedded processors from 2.2 Watts in 2004 to 3.0 Watts in 2018; corresponding numbers for the high-performance desktop variant are 156 Watts and 300 Watts. So the high-performance microprocessors consume 2 orders
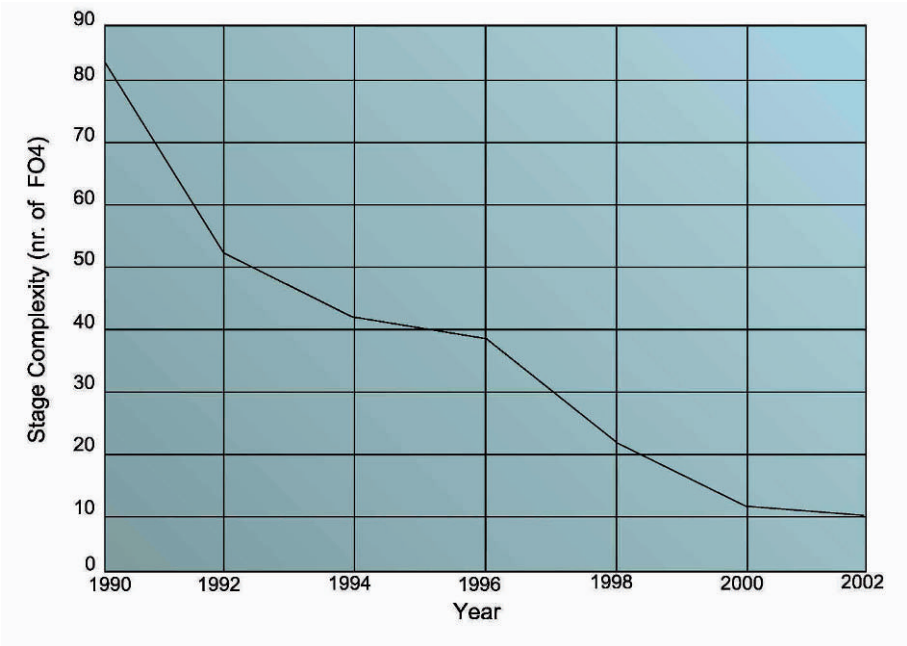
Figure 4.4. The decreasing stage size.

of magnitude more power and will increase their power consumption twice as much compared to the embedded microprocessors in portable devices.

Applications, whether they are communication/Signal Processing algorithms or multi-media streaming algorithms, both have real-time constraints and in this respect performance is the primary design goal and power consumption is a close second and area which improves quadratically with technology scaling a distant third. The easiest way to improve performance is to increase the clock frequency, which implies taking less time to do the unit operation. To take less time for the same unit operation, one needs to improve the propagation delay. And since the interconnect delay scales negatively with technology and as a component of the total propagation delay it is starting to dominate compared to the device propagation delay which scales linearly with technology, there are limits to how much performance improvement can be achieved by scaling up frequency as the technology scales. The other option to improve performance by increasing frequency is to reduce the size of the unit operation implemented in each clock cycle. This is achieved by pipelining and is the first architectural evolution to improve the computational power. The size of unit operation or the length of pipeline segment has been shrinking steadily since 1990 as shown in Figure 4.4 for seven generations of Intel processors. The size of pipeline

segment is in terms of numbers of fan-out-of-four(FO4) inverters; where FO4 being an atomic logic operation. Increasing clock frequency by reducing the logic depth has reached a stage where the law of diminishing returns kicks in. This is elegantly explained by Flynn et. al. in [5] and summarised here.

Let T be the execution time of an instruction which is segmented into S pipelined stages and each pipelined stage has an overhead C associated with Clocking. The microprocessor completes one instruction per cycle but when disruptions happen with a frequency b, S-1 instructions are invalidated. Dubey and Flynn have derived the throughput G and the optimal S_{opt} obtained by differentiating G with respect to S.

$$G = \frac{1}{1 + (S - 1)b} \cdot \frac{1}{T/S + C} \tag{4.1}$$

$$S_{opt} = \sqrt{\frac{(1 - b)}{bC}T} \tag{4.2}$$

Increasing S, reduces the T/S factory in Eq4.1 and thereby increases the throughput, assuming that b is sufficiently small. Beyond a point, however, increasing S starts to hurt because in spite of small b, the $(S\text{-}1)b$ factor starts to dominate and will cause G to dip. So to improve G by increasing S, we can a) increase T but that would reduce the degree of overlap between subsequent executions and will negatively impact the performance, b) reduce the clock overhead which is hard to achieve as technology scales because of the large skew and c) reduce the disruption frequency b, which is again hard to achieve, because it is intrinsically dependent on the nature of logic and input data. Another negative consequence of reducing pipeline stage size is that the pipeline depth increases. As especially control processors encounter many hazards during execution (about 1 in 4 or 5 instructions is a branch in control code), deep pipelines are a headache in processors.

To move beyond the limits of improving performance by logic level pipelining, microprocessor architects have exploited inherent parallelism in programs by providing hardware support for it as the next step in architectural evolution for improving the performance. The hardware support comes in form of multiple instruction units that are capable of executing instructions in parallel, provided of course there is no data dependency among them. Detecting lack of data dependency comes in two variants. For traditional DSP algorithms, where the lack of data dependency can be statically analysed, compiled with relative ease, compilers statically schedule non data-dependent operations onto the parallel instruction units in a class of processors called VLIW. General purpose CPUs that execute a wide range of programs, many of them relatively more control intensive compared to DSP algorithms, would loose many opportunities of exploiting parallelism by statically scheduling data independent operations. Consequently, hardware support is provided to dynamically detect

and schedule data independent operations onto parallel instruction units in a
class of processors called superscalar. However, like in the case of increasing
the pipelining depth by reducing the depth of logic, increasing instruction level
parallelism (ILP) cannot be monotonically increased to reap improvement in
performance and a similar trade-off is involved that has been analysed in [7]
and summarised here. Consider a processor capable of issuing N instructions
in a cycle. If O is the overhead of increasing the instruction width by one and
d is the average latency of instruction, then the instructions per cycle (IPC) as
presented in [5] and the optimal value of N by differentiating IPC with respect
to N are shown as Eq4.3 and Eq4.4. The upper bound on ILP is of course the
intrinsic parallelism in the program. VLIW processors have the advantage of
lower overhead O, whereas superscalar processors emphasise minimising the
disruption and thus reducing d.

$$IPC = \frac{N}{1 + N \cdot d \cdot (1 + O \cdot N)} \tag{4.3}$$

$$N_{opt} = \sqrt{\frac{1}{d \cdot O}} \tag{4.4}$$

The next architectural innovation that the embedded processors have made
is accelerating the software by delegating some of the compute intensive, fre-
quently executed fragments of code to hardware and/or adding special instruc-
tions to execute a certain class of algorithms. For instance ARM11 family of
high-end embedded processors have SIMD vectorisation instructions to speed
up multimedia instructions. Tensilica offers a tool suite to analyse the appli-
cation program to identify a) commonly occurring computational expressions
and fuse them into a single operator instruction, b) need for SIMD vectorisation
and c) the need for flexible instruction length as extensions to the base instruc-
tion set. This has the advantage of flexibly offering the additional instruction
on as per need basis. All embedded processors offer the possibility of adding
special purpose co-processors that can be invoked and controlled by software.
The extreme case of this approach of accelerating software by implementing
parts of algorithm in hardware is to implement the complete algorithm as an
ASIC. Dedicated hardware is the most cost-effective way of implementing logic
as its datapath is customised to the need of executing just one algorithm and
the overhead of fetching instructions, branch prediction, caching etc. are all
avoided. Dedicated hardware however has no flexibility of changing or hosting
another algorithm. This lack of flexibility and large NRE cost associated with
implementing ASICs is the reason why many architects are exploring the mid-
dle ground of retaining the flexibility while trying to achieve the performance
and energy efficiency of dedicated hardware. This quest has fueled the area
of reconfigurable processors and the Pleiades architecture at UC Berkely is a
prime example of it. More recently Coresonics (see [8]) has come up with a
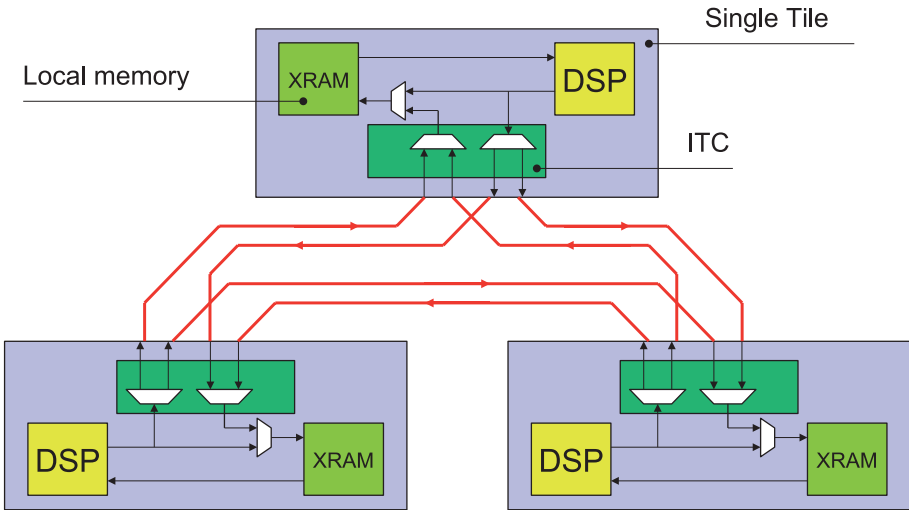
Figure 4.5. Sea of DSP message passing platform from Philips Semiconductors.

similar technology. Jan Rabaey(see [9]) presents data that shows that the energy efficiency of dedicated hardware can be three orders of magnitude more than that of a vanilla embedded microprocessor.

Hardware acceleration can speed up an algorithm or a class of algorithms, but when, as the present day embedded SOCs need, to host several high-performance concurrent and largely independent algorithms, hardware acceleration is not adequate. To address such situation, most embedded SOCs these days have multiple processors and/or dedicated hardware macros for algorithms that have stabilised in their specification, that are used very frequently and need ultra low power consumption. MP3 playback is a good example of it but not the only one. Other plausible use cases are for radio modems for 3G and WLAN standards. These multi-processor SOC architectures come in two variants and the two can be combined. The multiprocessor architecture can be built using message passing or shared memory concepts or a combination thereof.

Message passing is typically deployed in scenarios, where a) communication between processors are relatively well defined and b) the logic distributed among the multiple processors is tightly coupled. In other words, communication happens in well defined patterns and the producer consumer relationship and rates are well matched. This minimises the need for buffering is minimal because message passing architectures typically have distributed computational capability with small buffers tightly coupled to light weight processors. A good exponent of message passing based architecture in the commercial world is the
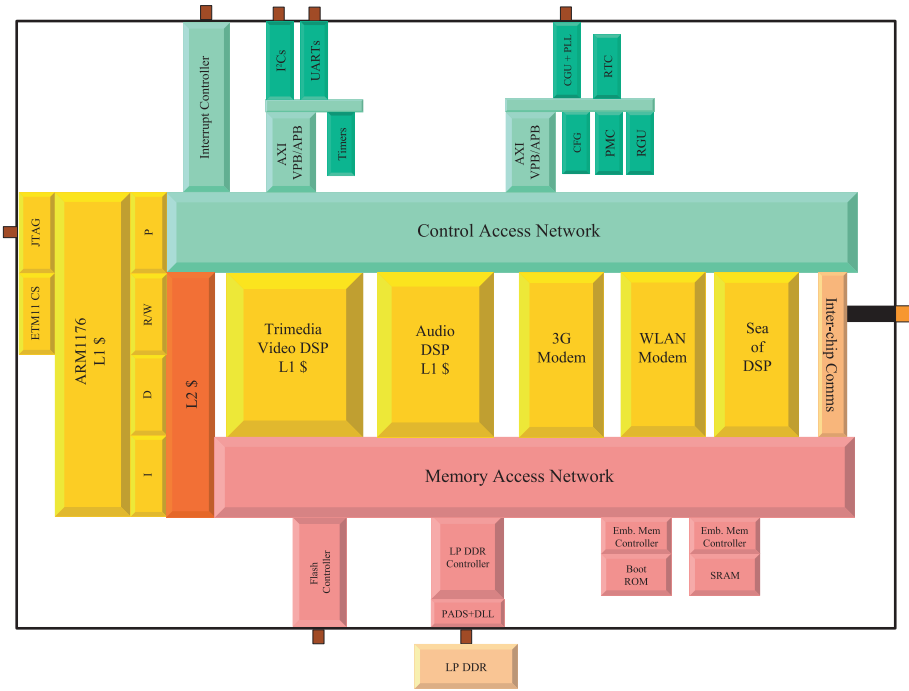
Figure 4.6. Multi-processor architecture shared memory architecture from Philips Semiconductors.

Sea of DSP (SOD) Platform from Philips Semiconductors (see [10], [11]). The key concept behind SOD is to build a distributed and scalable processing platform from small DSPs with local memory. DSPs execute small well defined atomic functions, mainly audio related today. These functions are small to fit onto a single DSP and all the data it needs resides locally, however it produces data that can be written remotely via the Inter Tile Communication(ITC) module, i.e. the DSPs operate in local read/write and remote write mode. A wide range of DSPs can be accommodated, only the tiling pattern remains same making it easy to build an arbitrarily large system. The entire Sea of DSP platform is controlled and configured by a system processor like ARM7 and can be part of a larger SOC as a macro.

Many high-end embedded SOC instances involve implementing a set of high-performance functions that are loosely connected and require almost arbitrary communication pattern among the resources implementing these complex functions. Moreover, these functions work with very large datasets and loose coupling requires large amount of buffering. Message passing, as presented above poorly matches these requirements and requires SOC architects to use

shared memory based inter processor communication. Shared memory architectures are characterised by a) large shared memory capacity as the name implies and b) high performance shared chip level interconnect and a system level processor that initiates, configures and controls the SOC.

A multi-processor SOC platform from Philips Semiconductors shown in Figure 4.6 is a good example of shared memory SOC architecture. The salient features of this architecture are:

- ARM1176 is the system controller with four high bandwidth AXI interconnects (AXI represents the present embedded interconnect protocol used in ARM Ltd, and replaces the older generation protocol AHB); the I and D ports couple the Instruction and Data L1 caches and the RW port connects the Tightly Coupled Memory(TCM) to the L2 and L3 Memories and the fourth port, the P(peripheral) port, is dedicated to controlling and configuring peripherals and other sub-systems.

- The architecture has two AXI interconnects. One for memory access and the other for control and peripheral access. The latter is used to access low bandwidth peripherals like UARTs, I^{2}C etc. but also to control and configure peripherals and sub-systems like Video DSP, Audio DSP and modems. The memory access network obviously has the higher bandwidth of the two networks and connects the ARM1176's three memory access ports and similar memory access ports of Video, Audio DSPs and other sub-systems to the memory hierarchy.

- Memory hierarchy consists of L1 memories encapsulated in the processors and the modem sub-systems, and Level L2 cache for the ARM1176 that has a greater need for a high bandwidth memory access. Further, there is onchip scratch pad SRAM memory and boot ROM. The main system memory is the external LP DDR memory and an external flash is used as a non volatile system storage.

This shared memory multi-processor architecture does provide the flexibility of implementing the loosely coupled applications, in need of large buffering and arbitrary interconnection among applications via memory. But this flexibility also comes with a penalty. Package and cost considerations often restrict a single large system memory that is shared among many subsystems. In spite of L2 caching for ARM1176, the architecture can potentially have a bottleneck in all sub-systems trying to access the external system memory.

2.2   Trends in Memory Usage in SOC

The amount of embedded memory in SOC is rapidly increasing. According to the SIA roadmap it has increased from 20% in 1999 to 70% today and will exceed 95% in a decade as shown in Figure 4.6.
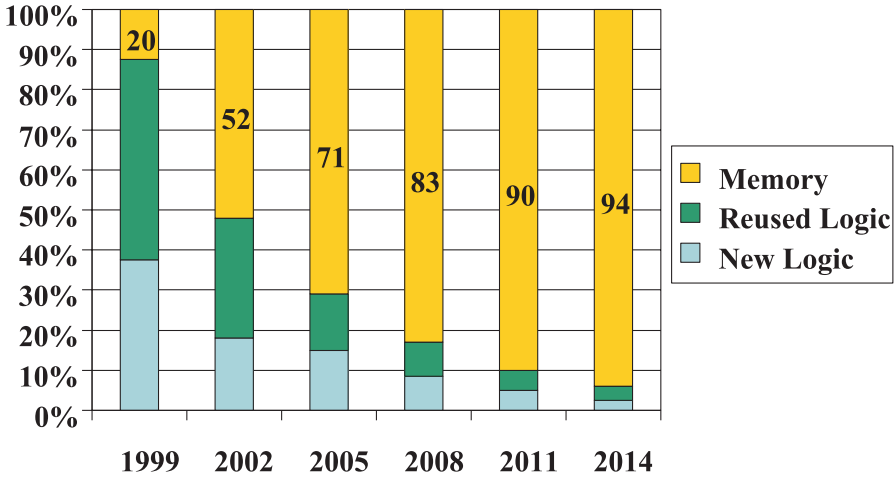
Figure 4.7.   Embedded memory content in SOC is increasing rapidly.

The increase in memory content comes from two sources primary and secondary. The primary source is rapid increase in the number and complexity of applications hosted by SOCs. These applications have not only large code size as suggested by their complexity but the nature of these high resolution and bandwidth algorithms works with large volume of streaming data that requires not just large memories but also a high bandwidth access to them. The secondary source comes from the need to cache instruction and data to boost performance by avoiding disruptions or minimising the penalty if a disruption happens. Until 2003, caching in SOC was mostly L1 instruction and data caches for individual processors. Since 2004, with the advent of high end processors like ARM11 that are clocked at 400 MHz or more in DSM technologies, L2 cache for individual processors or a system L2 cache is becoming increasingly common in high-end SOCs. Ad Siereveld, Memory Architect at Philips Semiconductors, has proposed a conjecture: *"In SOCs, the level of caching seems to follow the prevalent DDR generation".* There is some evidence to justify this conjecture: pre 2003 SOCs had L1 caches when DDR or SDR was the prevalent technology, since 2004 when DDR2 variants started getting adopted L2 caches are becoming common in SOCs and quite possibly L3 caches, which are already being architected in high end desktop processors, will become a feature in SOCs in 2008, when according to the JEDEX roadmap DDR3 is planned to be introduced.

Another noticeable trend as shown in Figure 4.8 is the move from static to dynamic memories to fulfill the need for large and fast volatile memories. This
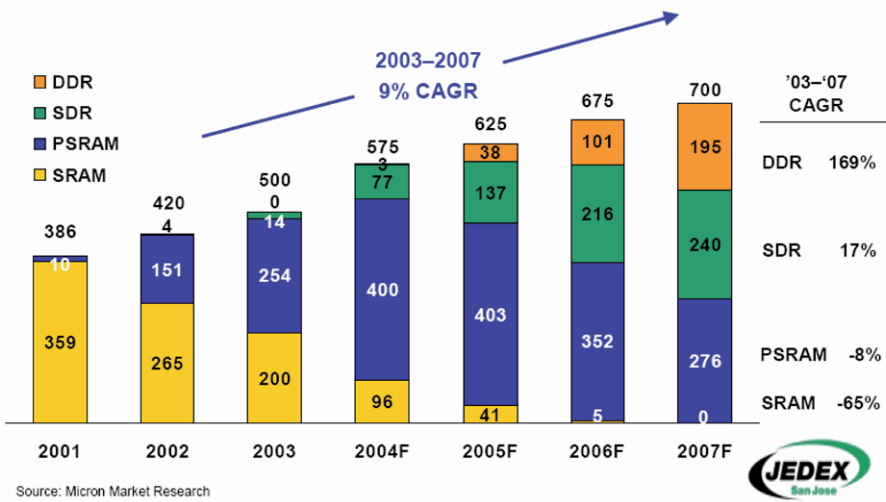
Figure 4.8. Volatile Memory is shifting from static to dynamic memories. (Source [12]).

is not surprising because SDRAMs have always enjoyed the superior density compared to SRAMs but now with DDR variants the speed issue is becoming manageable. In fact, there is some evidence from Hynix(see [13]) that memory speed trend is increasing exponentially while the CPU speed trend seems to be flattening. Among the DDR variants, the LPDDR (the Low Power DDR) is the most interesting for the battery powered SOCs. LPDDR with its features like temperature sensitive refresh rate, partial array refresh and extremely low self-refresh current is getting rapidly adopted by power constrained SOC designs.

2.3   Trends in Interconnect Architecture

Interconnects are the backbones of the SOC architectures. As SOCs have evolved to have multiple processors, the SOC architecture has become interconnect centric from the earlier processor centric. By being interconnect centric, SOC architectures have benefited from availability of sub-systems, specialized processors and peripherals that adhere to a certain interconnect standard. This simplifies rapid composition of complex SOC architectures, though more needs to be done in this direction as detailed later in this section.

Interconnects have evolved from being a replacement for board level interconnect on silicon to an architecture that is more in tune with the high performance requirements of today and one that exploits the relative abundance of routing resource on silicon compared to board level. Figure 4.10 shows evolution of interconnect schemes over three generations. The first scheme, Figure
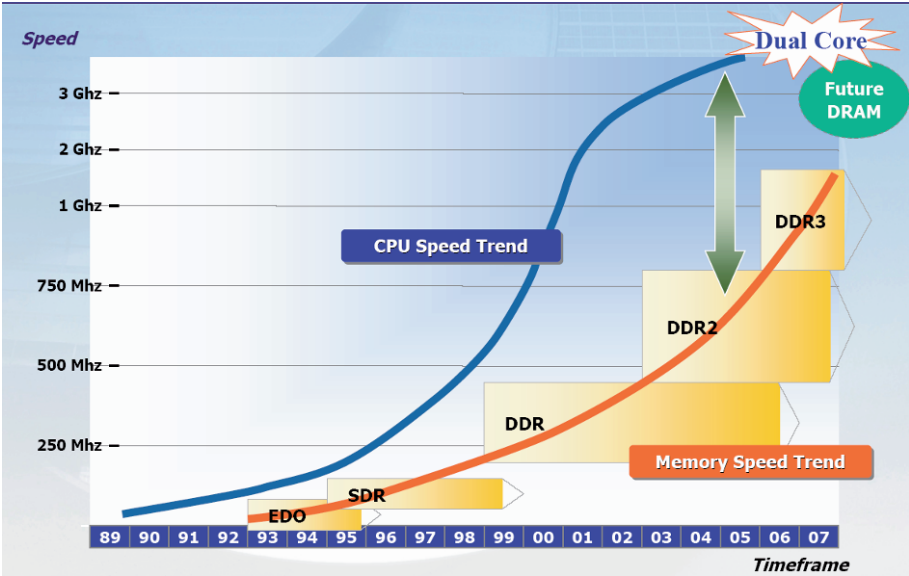
Figure 4.9.   CPU vs Memory Speed Trends. Source: Hynix (Source [13]).
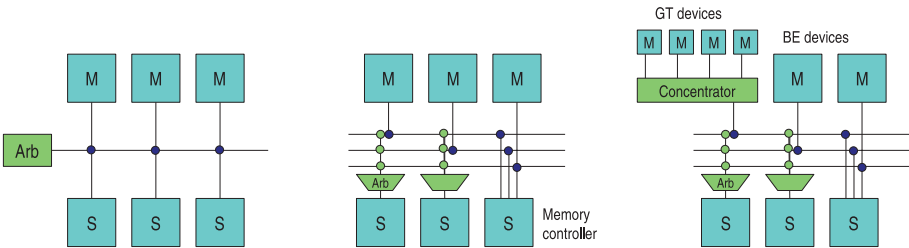
Figure 4.10.   Evolution of Interconnect Schemes.

4.10(a), is a multi-master arbitrated scheme, for which the bus interconnect is the shared resource. The classic AHB is representative of this scheme.

To overcome the shared bus bottleneck, the next generation scheme evolved to have one bus layer per master with slave side arbitration as shown in Figure 4.10(b). Multi-layer AHB is representative of this generation. This scheme moves the bottleneck from bus to slave interface level allowing concurrent transactions between different masters and slaves. As in many SOC it is a single high performance (SDRAM) memory controller that aggregates most of the communication bandwidth from the masters, this memory controller is typically equipped with multiple slave ports.
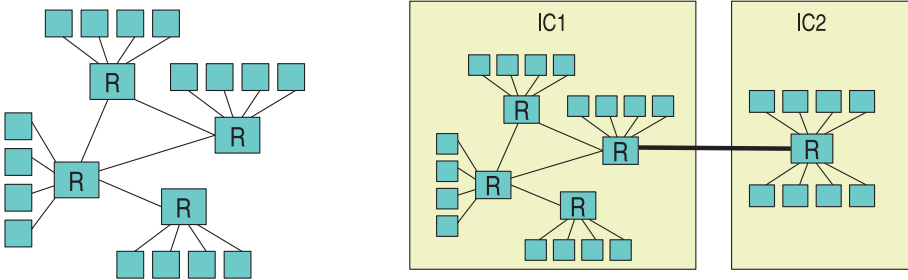
Figure 4.11. Evolution of NOC as intra-chip interconnect as well as an interchip interconect.

As the complexity of SOCs continues to increase, the number of masters with widely varying demands on interconnect bandwidth has increased as well. Applying the interconnect scheme shown in Figure 4.10(b) would lead to an unjustified increase in routing resources. Consequently, this leads to further evolution shown in Figure 4.10(c) where concentrators allow many masters to share one bus layer. Concentrators spread available bandwidth across masters according to the Quality of Service requirements of the individual masters (throughput, latency). Differentiation between guaranteed throughput (GT) and best effort (BE) traffic classes can be efficiently performed through separate bus layers. A related approach is the separation of control related traffic from memory directed traffic by means of implementing separate dedicated buses. Multi-Layer AXI is representative of this interconnect generation.

Looking forward, SOC evolution will continue to pose a challenge on interconnect technology, at SOC level as well as between chips (e.g. System in Package). Key issues to be addressed are improved wire utilization, flexibility in supporting different communication patterns, and flexibility in system composition. Packet routing Networks on Chip (NOC) have been studied in the academic world for a long time (see [14]) and are now at the verge of being applied at industry level (Philips Aethereal [15]). Very likely, NOC will emerge as the chip level backbone, whereas sub-systems will continue to be composed using conventional interconnect schemes. NOCs are expected to extend across chip boundaries. See Figure 4.11.

Besides evolving in a structural sense, interconnects have also evolved significantly in their protocols. The previous generation protocols, represented by AHB, supported one transaction at a time only, had inefficient burst models (redundant address transfers during bursts) and they coupled command with data flows, as shown in Figure 4.12(a). These features cause a severe inefficiency whenever traffic needs to be merged in a concentrator or in front of a slave.

This shortcoming has been overcome in the present generation protocols like AXI as shown in Figure 4.12(b). AXI supports multiple concurrent transactions,
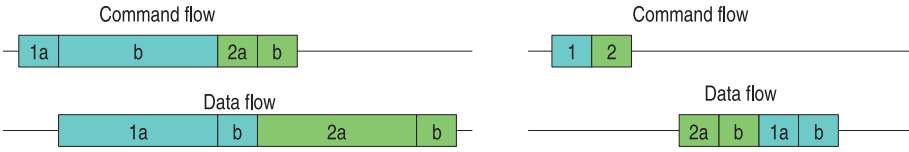
Figure 4.12.   Evolution of Interconnect Protocols.
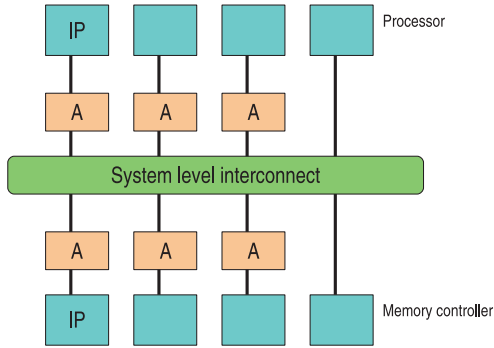
Figure 4.13.   Abstracting IP Functions from System.

has a burst model that avoids redundant address transfers, de-couples command from data flow, and it allows out-of order completion of transactions in different threads.

The described variety and evolution of system interconnect poses a challenge on the applicability of IP functions across systems and over time. To address this issue, companies have started to de-couple (abstract) IP functions from the system bus interconnect by means of adapter functions, as shown in Figure 4.13. IP cores communicate in an IP natural manner over their interfaces (i.e. using a data granularity, rate and width that reflects the actual data processing by the IP). Adapters then translate this IP natural communication into communication that is optimized for a particular bus/network and memory system, and they provide read and write buffers to hide system latencies from the IP.

The described approach is well applicable for all IP except processor and memory controllers. For performance reasons, the latter functions are typically attached directly to a system bus.
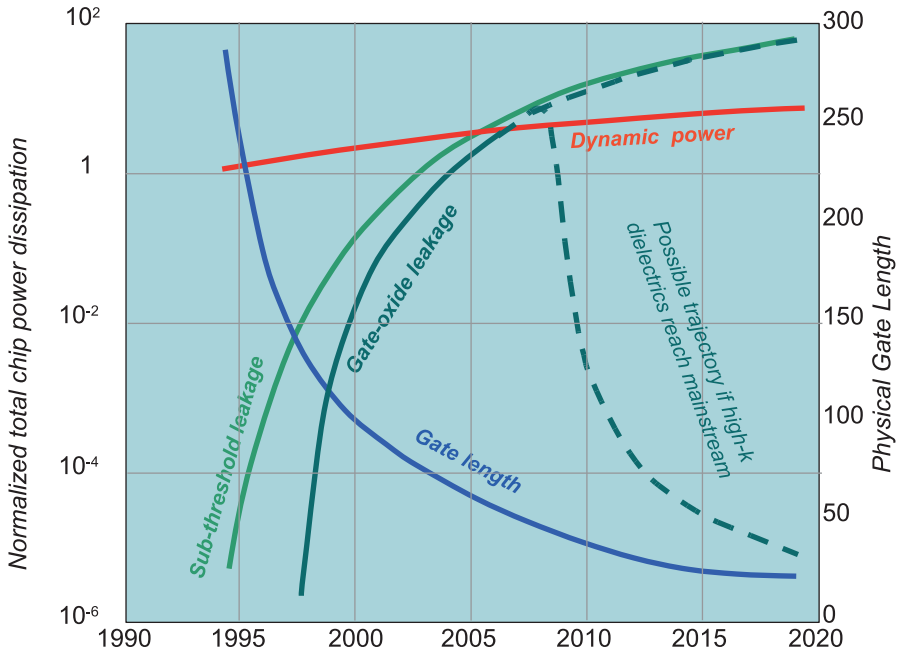
Figure 4.14. Trends in power consumption components. Source Nam Sung Kim et. al. [16].

2.4   Trends in Power Management

The component of power consumption, that for a long time was considered negligible, has with DSM geometries come to the forefront to the extent that warranted a title "Leakage Current: Moore's Law Meets Static Power"([16]) of a cover feature in IEEE Computer. The trend based on ITRS 2001 and adjusted to 2002 data reported in [16] is shown in Figure 4.14. It shows that the Dynamic Power Consumption, which was the dominant source of power consumption until 2000 is flattening out, whereas the two sub components of static power consumption, the sub threshold and the gate oxide leakage are exponentially rising and around 2005-2006 timeframe will overtake the dynamic power consumption. These trends are having a dramatic effect on the architectural solutions being adopted by the SOC community and constitute the architectural trends in SOC design that are aimed at bridging the power efficiency gap discussed in section 1.

Dynamic Power Consumption continues to be substantial and motivates architectural innovation. Two relations that define the policies to contain the dynamic power consumption are shown in Eq4.5 and Eq4.6. where f is the operating frequency, V is the supply voltage, V_t is the threshold voltage, C is

the total capacitive load and A is the fraction of the gates that switch and α is an experimentally defined constant that is 1.3 as reported in [16].

$$P_{dynamic} = ACV^2 f \qquad (4.5)$$

$$f = \frac{(V - V_t)^{\alpha}}{V} \qquad (4.6)$$

Clock gating is one of the most well established and practised dynamic power reduction methodologies and is automated at logic level by most logic synthesis tools. Clock gating is also manually controlled at block and subsystem level, where some control logic detects that a block or a sub-system is no longer active and turns off the clock. Dynamic power consumption reduces quadratically with supply voltage as suggested by Eq4.5. Until recently, this happened as technology scaled but no architectural innovation was made to exploit this relationship. As mentioned earlier, supply voltage has not scaled linearly with technology for reasons of standard and to maintain performance. The exponent α in Eq4.6 makes the operating frequency and thereby performance reduce exponentially with reduction in supply voltage. The computational load in multi-processor SOCs vary a lot over time. In other words, the operating frequency does not need to be the peak operating frequency all the time and by controlling the frequency based on computational load, the supply voltage and with it dynamic power consumption can be reduced as well. This scheme is called dynamic voltage frequency scaling(DVFS) and has been subject of intense research in last few years and is being adopted by industry.

Like DVFS for dynamic power consumption, architectural innovations have come up for static power consumption that are also based on the physics of sub-threshold and gate-oxide leakage. Chandrakasan et. al in [17] presents two relationships for the two leakage currents as shown in Eq4.7 and Eq4.8. Where K_1, K_2, n and α are experimentally determined, W is the gate width, T_{ox} is the gate oxide thickness and V_{θ} is the thermal voltage. These equations suggest mechanisms for reducing the two leakage currents. For I_{sub}, we can either reduce the supply voltage V to zero, thereby loosing the state but reaping the benefits of also reducing the I_{sub} to zero. The other alternative is to retain state but accept a lower performance by increasing the V_t, the negative exponent in Eq4.7. Both these mechanisms have been adopted to contain I_{sub} in practice. To reduce I_{ox}, Eq4.8 suggests that we should increase T_{ox}, whereas it instead scales down with the technology to avoid short channel effects. For this reason, the pursuit is for a high-K dielectric gate insulator as the solution and once that is in place the I_{ox} is expected to decline sharply as shown in Figure 4.14. Gate width is a common factor in both the leakage currents and reducing it is a design time option. If dynamic power is the dominant power consumption component, implementing logic in parallel is favoured compared to using the pipelined style of implementation to reduce the amount of logic that switches. However, if the
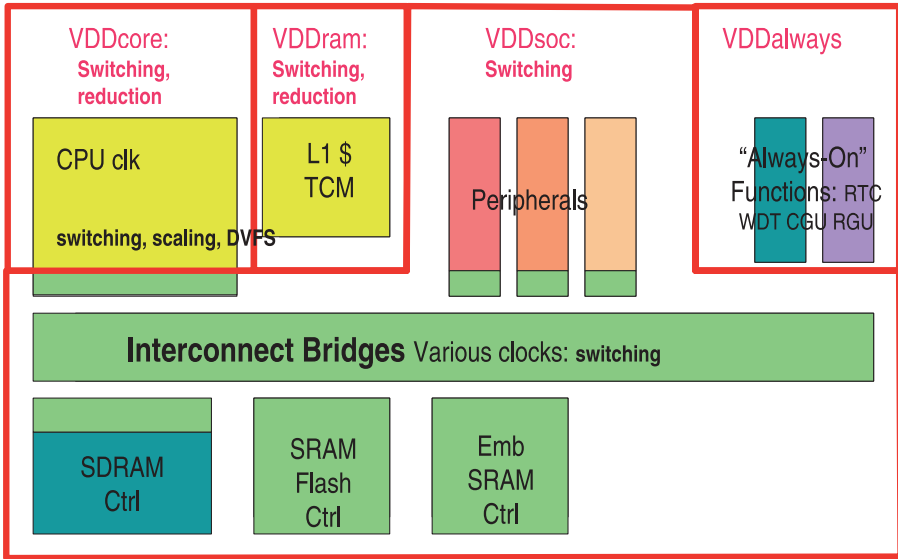
Figure 4.15. Multiple Voltage Domains.

static power consumption becomes the dominant component, pipelined style has the advantage compared to the parallel implementation because the pipelined style would reduce the gate width *W*. Other factors remaining constant, the total leakage current is proportional to the total gate width of a design.

$$I_{sub} = K_1 W e^{\frac{-V_t}{nV_\theta}} \left(1 - e^{\frac{-V}{V_\theta}} \right) \tag{4.7}$$

$$I_{ox} = K_2 W \left(\frac{V}{T_{ox}} \right)^2 e^{\frac{-\alpha T_{ox}}{V}} \tag{4.8}$$

Having discussed the conceptual basis for reducing dynamic as well as static power consumption, we present in an industrially realistic example of a power management architecture taken from a mobile application processor as shown earlier in Figure 4.6. Key elements are an ARM1176 processor, peripherals, on-chip interconnect, embedded memory and external memory controllers. The design is partitioned into power management domains for which VDD and clock supplies can be individually controlled. A further - very small - domain (VDDalways) remains always powered and clocked. It contains the power management and clock generation infrastructure.

For lowering dynamic power consumption, clock switching is applied as the baseline technique across all domains. Clock gating is performed transparent

to power management by synthesis tool created logic. For higher effectiveness, several components provide operating modes to stop clock supply under explicit control of power management. E.g. for this purpose the CPU provides a STANDBY mode that is invoked through software whenever the processor has no more computational tasks to perform, and that can be terminated by an enabled interrupt, when operation needs to be resumed. Clock switching can be performed instantaneously, hardly impacting the realtime behaviour of the system.

Assuming high dynamism of computational load, DVFS is the ideal dynamic power reduction technique for processors. DVFS is therefore applied for the CPU. The CPU can operate at several performance levels, each characterized by a discrete f, V operating point with V meeting the performance requirements (f) under worst case process and temperature conditions. To further improve effectiveness of DVFS, adaptive voltage scaling can be performed by letting the voltage supply operate in a closed regulation loop with an on-chip silicon performance monitor. This technique allows further reduction of the voltage level to what is required at the actual process and temperature conditions (typically being more relaxed than the worst case). It should be noted that DVFS is an expensive technique and therefore in practice restricted to few high power consumers in the system. Another note is that DVFS requires a careful physical implementation and control policy choice in order not to impact system performance: communication across voltage islands may introduce additional latencies, and a power management policy selecting a too low performance operating point may degrade responsiveness of the system or even lead to the miss of real-time deadlines.

For lowering static power consumption, voltage switching and voltage reduction techniques are applied by power management whenever components are inactive for a longer period of time (e.g. tens of milliseconds and above). In the example system, embedded switches perform switching of VDD supplies that are shared by multiple components; for other VDD domains switching is performed outside the SoC directly at the VDD source. VDD can be individually switched-off for the CPU core, RAM and SoC domains. Power management controls voltage switching through SHUTDOWN modes.

While voltage switching is perfect to reduce static power consumption to zero, it leads to total loss of state. State that needs to be preserved across SHUTDOWN mode must therefore be saved and restored using memory in a domain that remains powered. The example design therefore provides a DORMANT operating mode that maintains the state of CPU memories while the CPU core is powered off. In this mode, VDD supply to CPU memories is reduced to a minimum level that ensures state retention and minimal static power dissipation.

There are several attention points for applying the voltage switching and reduction techniques. Voltage islands that remain powered must be isolated from non-powered domains by signal clamping. An embedded VDD switch must be designed such that the switch itself does not lead to increased leakage, and the voltage drop over the switch is minimized at acceptable size (cost) of the switch. Real-time behaviour of the system may be affected by the latency required for VDD settling, and a power management policy should trade-off energy-saved by VDD switching against energy-consumed by a state save/restore process.

3.   Conclusion

To keep pace with demands of applications on performance and power consumption, SOC architectural components continue to evolve to bridge the gap between raw improvement due to technology scaling and what is required by the application. Embedded processors have evolved to the point that architecturally there is not much room for improvement in single processor architecture and thus they are evolving towards multi-processor configurations. The memory content in SOC designs is increasing exponentially. This is happening for two reasons: the first being the increasing complexity of applications that increases the need for both program and data memory, and the second being the drive to improve performance by reducing the number of disruptions and/or to minimise the penalty of disruption. With technology scaling and increasing complexity of SOCs, the interconnect delay has started to dominate and has become the bottleneck on performance and a major contributor to power consumption. Interconnect schemes have evolved primarily to relieve the performance bottleneck. More recently, embedded interconnect protocols, like AXI from ARM Ltd ([18]) have optional signals for power management. The most significant innovation in the chip level interconnect is the NOC, which allows distribution of clock in the form of packets that can be routed to their destination by on chip routers that are programmable. NOCs have the potential to enhance the interconnect performance, increase the usability of wires, provide redundancy, are more amenable to advanced low power management policies and above all will be the enabler for reusability at sub-system and platform level. For power management, the dominance of leakage current has been the main driver for innovation. Whereas turning off clock was the primary weapon against dynamic power consumption, turning off power supply is the key strategy against static power consumption. This has led to chips being divided into multiple voltage domains to allow turning off of power supply of those domains that are not active. To contain the dynamic power which continues to be substantial, DVFS has emerged as the new method beyond clock gating at various levels. While power and performance have been the drivers of innovation in SOC architecture so far, the next phase of innovation is likely to come from the need to cope

with the large process variation encountered as we continue to scale down the process geometries.

4.   Acknowledgement

References

[1] Jan M. Rabaey. Silicon Architectures for Wireless Systems - Part 1. Tutorial Hot-Chips 2001.

[2] Sven Mattisson. *Radio Design in Nanometer Technologies*, chapter Cellular RF Requirements and Integration Trends. Springer, 2006.

[3] Ahmed Hemani. Charting the EDA Roadmap. *IEEE Circuits and Devices Magazine*, pages 5–10, Nov/Dec 2004.

[4] Mary Jane Irwin, Luca Benini, N. Vijaykrishnan, and Mahmut Kandemir. *Techniques for Designing Energy-Aware MPSOCs. Multiprocessor Systems on chips*, chapter 2, pages 21–48. Morgan Kaufmann Publishers. ISBN 0-12-385251-X.

[5] Michael J. Flynn, Patrick Hung, and Kevin W. Rudd. Deep-submicron Microprocessor Design Issues. *IEEE Micro*, Jul/Aug 1999.

[6] Manfred Schlett. Trends in embedded processor design. *IEEE Computer*, Aug 1988.

[7] P. Hung and M. J. Flynn. Optimum ILP for Superscalar and VLIW Processors. Tech. Report CSL-TR-99-783, Stanford University, Dept. of Electrical Eng, 1999.

[8] Dake Liu, Anders Nilsson, and Eric Tell. *Radio Design in Nanometer Technologies*, chapter Programmable baseband DSP processors. Springer, 2006.

[9] Jan Rabaey, A. Abnous, H. Zhang, M. Wan, V. George, and V. Prabhu. Reconfigurable Processors - The Road to Flexible Power-Aware Computing.

[10] Ron et al. Schiffelers. Epics7B - A lean and mean concept. *IEEE ISPC GSPX*, 2003.

[11] Harpreet et al. Bhullar. Serving digital radio and audio processing requirements with sea-of-dsps for automotive applications the Philips way. *IEEE ISPC GSPX*, 2004.

[12] Dean Klein. Memory Trends, Drivers, and Solutions. 15-16 Apr 2004. www.download.micron.com/pdf/presentations/jedex/memory_trends_micron_2004.pdf.

[13] Reza Raramarzi. Inc. High Speed Trends in Memory Market. 1 Apr 2005. Hynix Semiconductor. www.jedex.org/images/pdf/reza_hynix_keynote.pdf.

[14] A. Hemani, A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg, and D. Lindqvist. Network on chip: An architecture for billion transistor era. In *Proc. IEEE Norchip Conf.*, pages 166–173, 2000.

[15] S.G. Pestana, E. Rijpkema, A. Radulescu, K. Goossens, and O.P. Gangwal. Cost performance trade-offs in networks on chip: A simulation-based approach. In *Proc. Design Automation, and Test in Europe Conf.*, volume 2, pages 764–769, 2004.

[16] Nam Sung Kim et. al. Leakage Current: Moore's Law Meets Static Power. *IEEE Computer*, pages 68–75, Dec 2003.

[17] A. Chandrakasan, W. Bowhill, and F. Fox. Design of high-performance Microprocessor Circuits. *IEEE Press*, 2001.

[18] ARM Limited. AMBA 3 AXI Specification.

Chapter 5

PROGRAMMABLE BASEBAND PROCESSORS

Dake Liu, Anders Nilsson, and Eric Tell

1.   Introduction

A typical wireless communication system contains several signal processing steps. In addition to the radio front-end, radio systems commonly incorporate several digital components such as the digital baseband processor, the media access controller and the application processor. An overview of such a system is presented in Figure 5.1.
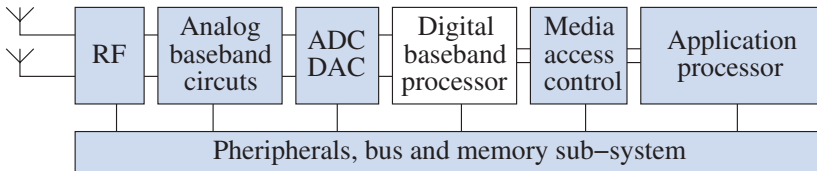
Figure 5.1.   Radio system overview.

In the transmitter, the baseband processor recieves data from the Media Access Control (MAC) processor and performs:

- Channel coding.

- Modulation.

- Symbol shaping.

before the data is sent to the radio front-end via a DAC. In the receiver the RF signal is first down-converted to an analog baseband signal which in turn is conditioned and filtered in the analog baseband circuitry. Then the signal is digitized by an ADC and fed to the digital baseband processor which performs:

- Filtering, synchronization, gain-control

- Demodulation, channel estimation and compensation.

- Forward error correction.

before the data is transfered to the MAC protocol layer.

The focus of this chapter is to give an introduction to programmable baseband processors for multi-mode radio systems and to highlight processing challenges which influence the design of such processors. Also a mapping of two popular modulation schemes, OFDM and (Wideband)CDMA onto a programmable processor is discussed.

2.   Baseband Processing Challenges

In this section some of the unique properties of baseband processing and some of the challenges faced in a wireless system are described. This section describes five general challenges which are common to most wireless systems:

- Multi-path propagation and fading. (Inter-symbol interference)

- High mobility.

- Frequency and timing offset.

- Noise and burst interference.

- Large dynamic range.

Handling these five issues impose a heavy computational load for the processor. Besides the above mentioned challenges, baseband processing in general also faces the challange of limited computing time and hard realtime requirements.

2.1   Multi-path Propagation

In a wireless system data are transported between a transmitter and a receiver through the air and are affected by the surrounding environment. One of the greatest challenges in wide-band radio links is the problem of multi-path propagation and inter-symbol interference. Multi-path propagation occurs when there is more than one propagation path from the transmitter to the receiver. Since all the delayed multi-path signal components will add in the receiver, inter-symbol interference will occur. Since the phases of the received signals

depend on the environment, some frequencies will add constructively and some destructively, thus destroying the original signal. Multi-path propagation is illustrated in Figure 5.2.
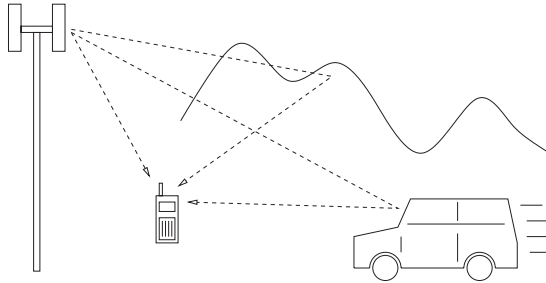
Figure 5.2. Multi-path propagation.

2.2 Timing and Frequency Offset

As the transmitter and the receiver in a wireless system will use different reference oscillators, a slight discrepancy will exist between the transmitter and the receiver carrier frequency and sample rate. Uncorrected, this difference will limit the useful data rate of a system. In addition, *doppler-spread* which is a frequency dependent frequency offset caused by mobility will further increase the frequency offset.

2.3 Mobility

Mobility in a wireless transmission causes several effects, both doppler-spread and rapid changes of the channel. The most demanding effect to manage is the rate at which the channel changes. If the mobility is low, e.g. when the channel can be assumed to be stationary for the duration of a complete symbol or data packet, the channel can be estimated by means of a preamble. However, if mobility is so high that the channel changes are significant during a symbol period, this phenomenon is called *fast fading*. Fast fading requires the processor to track and recalculate the channel estimate during reception of user payload data. Hence, it is not enough to rely on an initial channel estimation performed at the beginning of a packet or a frame.

2.4 Noise and Burst Interference

Noise and burst interference will degrade the signal arriving at the receiver in a wireless system. Both man-made noise and natural phenomenon such as lightning will cause signal degradation and possible bit errors. To increase the reliability of a wireless link *forward error correction* (FEC) techniques are

employed. In addition *interleaving* is often used to rearrange neighboring data-bits in order to even out bit-errors caused by burst interference or frequency selective fading. Popular FEC algorithms and codes are the Viterbi algorithm used for convolutional codes, Turbo codes or Reed-Solomon codes.

Dynamic Range

Another problem faced in wireless systems is the large dynamic range of received signals. Both fading and other equipment in the surroundings will increase the dynamic range of the signals arriving at the radio front-end. A dynamic range requirement of 60-100 dB is not uncommon. Since it is not practical to design systems with such large dynamic range, automatic gain control (AGC) circuits are used. This implies that the processor measures the received signal energy and adjusts the gain of the analog front-end components to normalize the energy received in the ADC. Since signals falling outside the useful range of the ADC cannot be used by the baseband processor, it is essential for the processor to continuously monitor the signal level and adjust the gain accordingly. Power consumption and system cost can be decreased by reducing the dynamic range of the ADC and DAC as well as the internal dynamic range of the number representation in the DSP processor. By using smart algorithms for gain-control, range margins in the processing chain can be decreased.

Processing Latency

Since baseband processing is a strict hard real-time procedure, all processing tasks must be completed on time. This imposes a heavy peak work-load for the processor during computationally demanding tasks such as channel decoding, channel estimation and gain control calculations. In a packet based system, the channel estimation, frequency error correction and gain control functions must be performed before any data can be received.

This may result in over-dimensioned hardware, since the hardware must be able to handle the peak work load, even though it may only occur less than one percent of the time. In such cases programmable DSPs have an advantage over fixed function hardware since the programmable DSP can reschedule its computing resources to make use of the available computing capacity all the time.

3.   Programmable Baseband Processors

Since baseband processing is computationally very heavy, baseband process-ing solutions have traditionally been implemented as fixed function hardware.

There are two major drawbacks of using non-programmable devices. The first is its low flexibility and short product life time. A fixed function product must be re-designed whenever there is a change in the product specification

whereas a programmable solution only needs a software update. The second is the excessive need for hardware resources. Designers seldom use hardware multiplexing techniques for digital baseband processing modules because of the added complexity and excessive verification time. If the module is not programmable, we cannot dynamically allocate computing resources to the respective algorithm, which implies that each function must be mapped to its own specific hardware.

3.1 Multi-mode Systems

As multi-mode radio terminals become popular, more attention must be paid to the design of the baseband processing hardware. A high-end cellular phone will support a number of standards such as: GSM/GPRS, EDGE, UMTS, WLAN, WiMAX, UWB, Bluetooth, GPS, and DVB-H. The classical way to design multi-mode systems is to integrate many separate baseband processing modules, each module covering one standard, in order to support multiple modes. One large drawback of using multiple non-programmable hardware modules is the large silicon area used and the lack of hardware reuse.

The trend is to utilize *programmable* baseband processors instead of fixed function hardware. Then several standards can be implemented with the same hardware, and the function can be changed by just running a different program [1-4].

In the following sections we will present baseband specific features of *Application Specific Instruction set Processors* (ASIP) and use the LeoCore DSP family from Coresonic [3] as an example.

3.2 Dynamic MIPS Allocation

By dynamically redistributing available resources, we can focus on either mobility management or high data rate. In Figure 5.3, the MIPS floor is limited by the top edge of the triangle. During severe fading we run advanced channel tracking and compensation algorithms to provide reliable communication. In good channel conditions more computing resources can be allocated to symbol processing tasks to increase the throughput of the system.

3.3 Hardware Multiplexing Trough Programmability

The concept of HW multiplexing is shown in Figure 5.4.

Most wireless communication uses modulation schemes which can be divided into three classes: OFDM, CDMA and single carrier modulation.

As illustrated by Figure 5.4 all these modulation schemes can be implemented on a DSP with the functionality shown in the figure. By carefully selecting the functional blocks maximum hardware reuse between different standards and modulation schemes can be achieved.
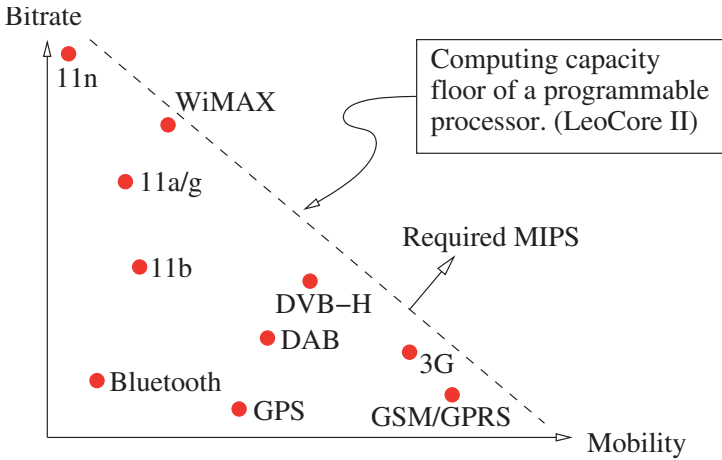
Figure 5.3.   Dynamic MIPS usage.

Figure 5.4.   Hardware multiplexing on LeoCore DSPs.

4.   OFDM and WCDMA Example

In this section we use OFDM and CDMA based standards to exemplify the requirements of a baseband processor.

4.1   Introduction to OFDM

Orthogonal Frequency Division Multiplexing (OFDM) is a method which transmits data simultaneously over several sub-carrier frequencies. The name comes from that fact that all subcarrier frequencies are mutualy orthogonal, Thereby signalling on one frequency is not visible on any other sub-carrier frequency. This orthogonallity is achieved in a nice way in implementation by collecting the symbols to be transmitted on each sub-carrier in the frequency

domain, and then simultaneously translating all of them into one time domain symbol using an inverse fast fourier transform (IFFT).

The advantage of OFDM is that each sub-carrier only occupies a narrow frequency band and hence can be considered to be subject to flat fading. Therefore a complex channel equalizer can be avoided. Instead the impact of the channel on each sub-carrier can be compensated by a simple multiplication in order to scale and rotate the constallation points to the correct position once the signal has been transfered back to the frequency domain (by way of an FFT) in the receiver.

To further reduce the impact of mulitpath propagation and intersymbol interference (ISI), a guard period is often created between OFDM symbols by adding a cyclic prefix (CP) to the beginning of each symbol. This is achieved by simply copying the end of the symbol and add it in front of the symbol. As long as the channel delay spread is shorter than the cyclic prefix, the effects of ISI are mitigated.

4.2   Job Overview

lready at this point it should be quite clear that efficient calculation of the FFT is vital for an OFDM baseband processor. In order to illustrate the amount of processing needed, Table 5.1 gives an overview of some well known radio standards using OFDM [5], [6], [7].

The last line of the table shows the approximte MIPS cost needed only for the FFT itself if the OFDM tranciever is implemented in a general DSP processor (with FFT adressing support). However, FFT is not the only demanding task in an OFDM tranciever. Other heavy jobs adding to the baseband processor requirements are synchronization, channel estimation and channel decoding.

Figure 5.5 shows typcial OFDM processing flows for packet detection/ synchronization/channel estimtation, for payload reception, and for transmission. Essentially all processing between mapping/demapping and ADC/DAC manipulates I/Q pairs represented as complex values. The remaining part of the baseband processing consists mainly of channel coding/decoding which tyically consists of bitmanipulation operations. Channel coding will not be discussed here since it is, with few exceptions, not OFDM-specific.

Table 5.2 shows some of the processing steps, the types of operations involved and approximate DSP MIPS cost for the 802.11a standard.

4.3   Hardware Considerations for Programmable OFDM Processing

FFT Acceleration

Since the FFT is the major corner stone of OFDM processing it may seem logical to employ a dedicated hardware accelerator block for FFT. Especially

Table 5.1.   FFT computation complexity for different OFDM standards.

standard	*802.11a/g*	*WiMax*	*DVB-H (4k mode)*
application	Wireless LAN	Wireless Access	Digital TV
max bitrate	54 Mbit/s	46.6 Mbit/s	32 Mbit/s
sample rate	20 MHz	13.8 MHz	9.1 MHz
FFT size	64	256	4096
symbol rate	250 kHz	43 kHz	2.2 kHz
with radix-2 FFT:			
processing	48 Mbf/s	44 Mbf/s	53 Mbf/s
mem bandwidth	288 Msample/s	265 Msample/s	319 Msample/s
memory size	320 samples	1536 samples	32672 samples
with radix-4 FFT:			
processing	12 Mbf/s	11 Mbf/s	13 Mbf/s
mem bandwidth	144 Msample/s	133 Msample/s	160 Msample/s
memory size	272 samples	1280 samples	26528 samples
equiv. DSP MIPS	480	440	530

since the implementation of such hardware has been widely studied and very efficent solutions exist, e.g. radix-2^2 implementations [8]. However our experience is that in a programmable solution it is usually more suitable to only accelerate FFT on instruction level by adding butterfly instructions together with bit-reversed/reverse-carry addressing support. There are two main resons for this:

Flexibility: Many fixed function FFT implementations tend to loose much of their advantage if multiple FFT sizes must be supported. With butterfly instructions and bit reversed adressing support one has full flexibility to efficently implement any size of FFT. As a bonus other types of transforms, such as cosine- or Walsh transforms can also be supported.

Hardware reuse: Even the most efficent FFT implementation will contain large hardware components such as (complex valued) multipliers and hence occupy a significant silicon area. If instead one uses dedicated instructions executing in the core data path/MAC unit these expensive hardware components can be reused by completely diffrent instructions and algorithms.

Table 5.2.   OFDM algorithm profiling

Function	Operations	MIPS
Receive/decimation filter	FIR/IIR filter: CMAC	1200
Packet detection	Autocorrelation: CMAC	320
Frequency offset estimation	Autocorrelation, complex argument calculation: CMAC, cordic algorithm	100
Frequency offset correction	rotor: table look-up, CMUL	480
Synchronization	Crosscorrelation in time domain: CMAC, absolute maximum *or* in frequency domain: FFT, CMUL, IFFT, absolute maximum	400
Channel estimation	Frequency domain correlation with known pilot symbol: FFT, CMUL	400^a
Channel equalization	One complex multiplication for each sub-carrier: CMUL	64^b
Demodulation	FFT	480

[a] MIPS saved my combining synchronization and channel estimation.
[b] Integrated with FFT.

This kind of hardware multiplexing in fact often means that a programmable solution in many cases can reach a smaller total silicon area than a corresponding fixed function solution.

4.4   Introduction to CDMA

Code Division Multiple Access (CDMA) is a multiple access scheme which allows concurrent transmission in the same spectrum by using orthogonal *spreading codes* for each communication channel. In this section the two CDMA based standards: Wideband CDMA (WCDMA) and High Speed Data Packet Access (HSDPA) are used as examples.

In a CDMA transmitter, binary data are mapped onto complex valued symbols which then are multiplicated (spread) with a code from a set of orthogonal codes. The length of the spreading code is called the *spreading factor* (SF). In the receiver data are recovered by calculating a dot-product (de-spread) between the received data and the assigned code. Since the spreading codes are selected from an ortogonal set of codes, the dot-product will be zero for all other codes

preamble   receive   transmit
processing   processing   processing

filtering	filtering	symbol shaping
AGC, packet detection	frequency offset compensation	
frequency offset estimation	remove CP	add CP
symbol sync.	FFT	IFFT
channel estimation	channel equalization	pilot insertion

complex−valued processing
- - - - - - - - - - - - - - - - demapping - - - - mapping - - - - - -
bit−based processing

deinterleaving	interleaving
error correction	channel coding
descramble	scramble

Figure 5.5.   OFDM processing flow.

except the assigned code. By varying the spreading factor, the system can trade data rate against SNR as a higher SF increase the energy per symbol.

A feature of WCDMA is the ability to scale the bandwidth to a particular user by assigning multiple spreading codes to that user. Using multiple codes is reffered to multi-code transmission. Multi-code transmission can also be used for *soft handover*, e.g. when the mobile station is handed over between two or

more base-stations. By using one or many codes from each of the involved base-stations, the mobile can be handed over without any interruption of the service. The WCDMA standard requires the mobile to manage up to 3 simultaneous codes on 6 base stations (18 codes).

4.5   Job Overview

The signal processing in WCDMA and HSDPA can be divided into *chip-rate* processing and *symbol-rate* processing. A *chip* is one complex element of the spreading code. Synchronization, channel estimation and channel equalization is performed in chip-rate, whereas additional channel equalization is performed in symbol rate.

Synchronization

The synchronization block in a WCDMA terminal is responsible of finding the start of a data frame and identifying base station parameters. This is ac-compished by correlating received data with a 256 chips long synchronization code. Furthermore are the result from the correlation used to identify a number of strong multi-path components. As illustrated in Table 5.3 the chip-rate of WCDMA/HSDPA is 3.84 Mchips/s. With an oversampling rate of four the multi-path components can be resolved with an accuracy of 20 meters. The main operation in this step is complex multiplication and accumulation (complex dot product).

Channel Equalization

Channel equalization is often performed in two steps in WCDMA. First a number of the strongest multi-path components are identified by using data from the synchronizer, the multi-path components are aligned in time and added constructivly (using maximum ratio combining). This is known as a *Rake*-receiver.

The performance of a rake-receiver is often adequate for WCDMA basic services. However, in HSDPA (which uses up to 16 QAM), additional equalization is necessary. In a HSDPA receiver, the resulting complex-valued symbols after de-spread is equalized by a second linear equalizer which uses traning symbols inserted in the middle of the data slot (midamble).

4.6   Hardware Considerations for a WCDMA Processor

As in the OFDM case, all chip and symbol related operations are performed on complex valued data, hence efficient complex computing is essential for a programmable baseband processor.

Also, as the processor operates on fairly short symbols with a high symbol rate, the loop overhead must be minimized. By employing wider execution

Table 5.3.   Computation complexity for WCDMA-FDD and HSDPA.

standard	*WCDMA-FDD*	*HSDPA*
application	3G Voice and data	Data packet access
spreading factor	4-512	4-512
modulation	QPSK	16 QAM
max bitrate	384 kbit/s	32 Mbit/s
chip rate	3.84 Mcps	3.84 Mcps
sample rate	15.36 MHz	15.36 MHz
symbol rate	7.5..960 kHz	7.5..960 kHz
synchronization:		
equivalent DSP MIPS	109	109
rake channel equalizer:		
# of fingers	18	18
mem bandwidth	73 Msample/s	73 Msample/s
memory size	160 samples	160 samples
equivalent DSP MIPS	384	384
symbol equalizer:		
equivalent DSP MIPS	-	256

units, processing efficiency can be improved. In WCDMA and HSDPA and
other CDMA systems the complex spreading codes have constant envelope,
which enables de-spread operations to be performed in a complex ALU instead
of entirely in a complex MAC unit. Addressing support for Rake-adressing is
also important. The addressing support is generally implemented as function
level accelerators in memory blocks.

5.   Multi-Standard Processor Design

In this section a processor architecture suitable for both OFDM, CDMA and
single carrier based standards is presented. To summarize the requirements
gathered from the OFDM and WCDMA example, the following points must be
considered. The processor must have:

1 Efficient instruction set suited for baseband processing. Use of both natively
   complex computing and integer computing.

Table 5.4. WCDMA algorithm profiling.

Function	Operations	MIPS
Receive/decimation filter	FIR/IIR filter: CMAC	322
Synchronization:	Crosscorrelation: CMAC	109
Frequency offset estimation	Autocorrelation, complex argument calculation: CMAC, cordic algorithm	160
Frequency offset correction	rotor: table look-up, CMUL	76
Channel estimation	Crosscorrelation: CMAC, absolute maximum	109[a]
Rake:	De-spread and MRC: CMAC	384
Time domain filter	FIR filter: CMAC	256

[a]MIPS saved by combining synchronization and channel estimation

2 Efficient hardware reuse trough instruction level acceleration.

3 Wide execution units to increase processing parallelism.

4 High memory bandwidth to support parallel execution.

5 Low overhead in processing.

6 Balance between configurable accelerators and execution units.

5.1 Complex Computing

A very large part of the processing, including FFTs, frequency/timing offset estimation, synchronization, and channel estimation all employ well known convolution based functions common in DSP processing. Such operations can typically be carried out efficiently by DSP processors thanks to complex multiply-accumulate (CMAC) units and optimized memory- and bus architectures and adressing modes. However in baseband processing essentially all these operations are complex valued. Therefore it is essential that also complex-valued operations can be carried out efficently. To reach the best efficency, complex computing should be supported throughout the architecture: by data paths and instruction set as well as by the memory archtitecture and data types.

5.2 LeoCore Processor Overview

The LeoCore DSP family from Coresonic is used as an example troughout this section. Some of the features and the architecture of the LeoCore DSPs are presented.

The LeoCore DSP consists of two main parts, one natively complex part which mainly operates on vectors of complex numbers and one natively integer part which operates on integers and single bits. The latter part is mainly used for forward error correction (FEC) and bit manipulation whereas the former part is used to extract soft data symbols that can be de-mapped into bits.

Furthermore, the memory system consists of memories which are connected to the execution units trough an on-chip network. The architecture is shown in Figure 5.6.
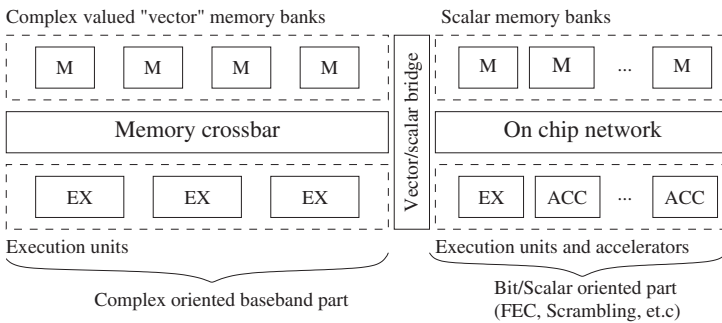
Figure 5.6. LeoCore basic architecture.

The on-chip network allows any memory to be connected to any execution unit. Execution units span the range from a DSP controller core, to multi-lane complex MAC and ALU SIMD data-paths. Accelerators are also attached to the network.

The architecture relies on the observation that most baseband processing tasks operate on a large set of complex-valued vectors (such as auto-correlation, dot-product, FFT and convolution). This allows us to optimize execution units to take advantage of this. The LeoCore architecture uses *vector instructions*, i.e. a single instruction that triggers a complete vector operation such as a complex 128 sample dot-product.

To support this kind of instructions, the execution units must be able to process large data chunks without any intervention from the processor core. This in turn requires the execution unit and memory sub-system to have automatic address generation and efficient load/store subsystems. As a response

to this, the base architecture utilizes de-centralized memories and memory addressing together with vector execution units.

5.3 Execution Units

To provide an efficient platform for multi-standard baseband processing, a baseband processor must provide several high-troughput execution units capable of executing complex tasks in an efficient manner.

The LeoCore family of DSPs utilize all complex valued execution units which range from CMAC units capable of executing a radix-4 FFT butterflies in one clock cycle to complex ALUs used by CDMA based standards.

5.4 Memory Subsystem

The amount of memory needed is often small in baseband processing, but the required memory bandwidth may be very large. As seen in Table 5.1 the FFT calculation alone may need a memory bandwith of several hundred Msample per second, averaged over the enitre symbol time (bear in mind that each sample consists of two values: the real part and the imaginary part). In practice the peak memory bandwith required may be several hunred bits per clock cycle for a processor running at a few hundred MHz. High memory bandwidth can be achieved in different ways - using wider memories, more memory banks, or multiport memories - resulting in different tradeoffs between flexibility and cost.

Baseband processing is chracterized by a predictable flow with few data dependencies and regular adressing, which means that flexible but expensive multiport memories often can be avoided.

The irregular (bit-reversed) adressing in FFT and Rake channel equalization could be considered an exception from this, however schemes exist which makes it possible to use only single port memories and still not cause memory access conflicts even if all inputs/outputs of each butterfly is read/written in parallel.

5.5 HW Acceleration

To further improve the computing efficiency of the processor, function level accelerators could be used. A function level accelerator is a configurable piece of hardware which performs a specific task without support from the processor core.

When deciding which functions to accelerate as function level accelerators the following must be must considered:

1 **MIPS cost.** A function with a high MIPS cost may have to be accelerated if the operation cannot be performed by a regular processor.

2 **Reuse.** A function that is performed regularly and is used by several radio standards is a good candidate for acceleration.

3 **Circuit area.** Acceleration of special functions is only justified if there can be considerable reduction of clock frequency or power compared to the extra area added by the accelerator.

An operation which fulfills one or more of the previous points is a good candidate for hardware acceleration.

5.6   Typical Accelerators

Front-end Acceleration

In most cases the recieved baseband signal will be subject to filtering/ decimation in the receiver before it is passed on to the kernel baseband processing. The required filter can be quite costly in terms of MIPS (again se Table 5.2). Since this function is needed in almost all radio standards and always runns as soon as the tranciever is in receive mode (receiving data or just waiting for data to appear on the radio channel) the filter is a suitable candidate for acceleration.

Several other functions may also be suitable to include in the same accelerator blocks. All these functions are very general and can be reused for many standards:

Resampling: E.g. a farrow structure can be used to receive standards with different sample rate using a fixed clock ADC clock *or* to compensate for sample frequency offset between transmitter and receiver.

Rotor: A rotor (essentially an NCO and a complex multiplier) can be used to compensate for freqency offset between transmitter and receiver. It can also be used for the final down conversion in a low-IF system.

Packet detector: The packet detector recognizes signal patterns that indicate the start of a frame. The baseband processor can then be shut down to save power, and be woken up by the packet detector when a valid radio frame arrives.

Shaping filter: During transmission this filter is used to shape the transmitted symbols. This filter is useful in full-duplex systems and can in certain situations be time-shared with the receive filter.

Forward Error Correction

Forward error correction functions are also a good candidate for acceleration since the three most common FEC algorithms are used in many different standards as shown i Table 5.5.

Table 5.5. FEC algorithms usage in common standards.

standard	Viterbi	Turbo	RS
IEEE 802.11a/g	x		
IEEE 802.11b			
WiMax	x	x	x
DVB-T	x		x
WCDMA	x	x	x
HSDPA	x	x	x

As the function of the FEC block is similar between different standards and the MIPS-cost is high [10] these blocks are often implemented as configurable accelerators.

6. Conclusion

Multi-standard baseband processing can be implemented efficently in a programmable baseband processor. The main features of the processor should be:

- Inherent support complex valued computing.

- Instruction level acceleration of FFT, convolution and similar kernel functions.

- Optimized memory architecture meeting the high bandwith- and real-time requirements but typically with a small total amount of memory.

In addition much of the channel coding tasks, as well as some general tasks close to the ADC/DAC interface are often suitable for function level acceleration. Selecting a good trade-off between programmability and function level acceleration ensures versatile yet efficient baseband processors.

References

[1] E. Tell, *Design of programmable baseband processors* Linköping Studies in Science and Technology, Thesis No. 1173, Linköping, Sweden, June 2005

[2] A. Nilsson, *Design of multi-standard baseband processors* Linköping Studies in Science and Technology, Thesis No. 1173, Linköping, Sweden, June 2005

[3] http://www.coresonic.com/ *Coresonic AB*.

[4] http://www.da.isy.liu.se/research/bp *Computer Engineering, Linköping University*.

[5] IEEE 802.11a, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications High-speed Physical Layer in the 5 GHz Band, 1999.

[6] IEEE Standard for local and metropolitan area networks, Part 16: Air Interface for Fixed Broadband Wireless Access Systems. WirelessMAN-OFDM

[7] ETSI EN 300 744, Digital Video Broadcasting. DVB-T/DVB-H

[8] H. Shousheng, M. Torkelsson, *Designing pipeline FFT processor for OFDM (de)modulation* Signals, Systems, and Electronics, 1998. ISSSE 98. 1998 URSI International Symposium on 29 Sept.-2 Oct. 1998 Page(s):257-262

[9] 3rd Generation Partnership Project; Physical channels and mapping of transport channels onto physical channels (FDD) 3GPP TS 25.211 V6.0.0

[10] A. Nilsson, E. Tell; *An accelerator structure for programmable multi-standard baseband processors.* Proc. of WNET2004, Banff, AB, Canada, July 2004.

Chapter 6

ANALOG-TO-DIGITAL CONVERSION
TECHNOLOGIES FOR SOFTWARE
DEFINED RADIOS

Ana Rusu and Mohammed Ismail

1.   Introduction

The trend in modern transceivers design is to move to software-defined radio (SDR) where reconfigurable hardware can support a variety of technologies (cellular, Bluetooth, WLAN, WiMAX, DVB-H, etc.) just by changing the software. In such a radio, coexistence and simultaneous use are challenges, but the major challenges on the mobile terminal side are the power dissipation, energy management, size and cost. Handsets that integrate cellular and Wireless Local Area Network (WLAN) have been already introduced to the market, but new emerging wireless technologies, such as 3G (Long Term Evolution or UTRAN-LTE) and Worldwide Interoperability for Microwave Access (WiMAX) have to be considered as well. Global Positioning Systems, FM radio and digital video broadcasting will also become part of future handsets. Together, these standards lead to challenging architectural requirements such as reconfigurability and programmability for multi-standard radio solutions that are both backup compatible and future proof. A key design consideration for mobile terminals is the analog-to-digital interface; the analog-to-digital converter (ADC) and digital-to-analog converter (DAC) can be the performance-limiting blocks in the receiver/transmitter chain. The requirements imposed by many wireless standards are often tougher to meet in the receiver chain than in the transmitter chain. Therefore the focus of this chapter is devoted to the receiver chain and ADC design challenges for SDR system solutions. This chapter discusses possibilities of implementing reconfigurable ADCs for Software Defined Radio handsets. The Software Defined Radio concept is introduced in Section

2 and commercial SDRs are presented in Section 3. The current and future radio architectures and their requirements for the ADC are introduced in Section 4. Section 5 focuses on the ADC challenges for SDR. Section 6 presents our approaches for implementing reconfigurable ADC architectures for SDR handsets. Finally, Section 7 concludes the chapter.

2.   Why Software Defined Radios?

Software Defined Radio is one of the emerging technologies being considered these days in many commercial embedded applications. John Mitola has introduced the Software Defined Radio (SDR) concept in the early 90's [1], and initially it was promoted only by the defense industry. The SDR concept was widely studied and promoted by the U.S. military in its multi-service Joint Tactical Radio System (JTRS). Now, SDR is moving into the commercial marketplace, where the number of players is rapidly increasing. New wireless services and standards are introduced and the consumers expect multifunctional mobile devices. So, the wireless industry is becoming more complex and the radio developers face special challenges in their design: multi-band antennas, multi-band multi-mode RF front-ends, multi-standard ADCs/DACs, and reconfigurable digital signal processors. Software Defined Radio allows a single wireless device to support multi-band and multi-mode radios previously available only through multiple devices. SDR is a radio with a generic hardware based on analog circuitry, under a flexible software architecture. Although the SDR concept has been around for many years, practical designs are only now becoming possible due to advances in many technologies, including: CMOS, analog-to-digital and digital-to-analog conversion, field-programmable gate arrays (FPGAs), ultra-fast data-transfer interfaces, powerful, cost-effective programmable digital signal processors (DSPs), and adaptive computing machines. SDR transceivers have several advantages over traditional radio transceivers. The most important advantage is flexibility. SDRs can be programmed and/or reconfigured on the fly. Users could initially configure their radios by downloading a personalized package of software features. Upgrades and reconfigurations could be done via simple internet access, and downloads could even be received over the air. The reconfigurable radios may offer a significant benefit to public service industries: emergencies such as flooding, volcano eruptions, train accidents, and tornados. Another important advantage is that SDR is an "agile" radio technology because it can also be configured to handle multiple communications protocols and technologies including Global System for Mobile Communications (GSM), Wideband Code Division Multiple Access (WCDMA), Bluetooth, WLAN, WiMAX and future standards. Finally, because of its modularity and flexible software architecture, SDR is a cost-effective solution for both manufacturers and end users.

3.   Commercial SDRs and SRs

The Software Defined Radio Technology includes reconfigurable radios, software defined radios (SDRs) and software radios (SRs). According to Vanu Inc. "Software Radio is a type of SDR that maximizes software reuse across platforms and hardware generations" [2]. The SDR [3] has not been achievable, until last year, due to the lack of ADCs capable of converting the RF signals directly to digital data. Vanu Software Radio^{TM}is the first commercially available Software Radio device where a single reusable hardware platform can support multiple wireless services and standards entirely in software. The system can support all of the GSM cellular base station functionality running on off-the-shelf Hewlett Packard ProLiant servers with an Analog Devices Corporation Digivance^{TM} RF subsystem. There is a lot of ongoing research and prototypes or commercial SDR products for the breakthrough technologies, as the one proposed by TechnoConcepts [4]. TechnoConcepts has produced a True Software Radio^{TM}(TSR) transceiver technology that replaces conventional analog circuitry with the combination of its proprietary delta-sigma converters and software based digital signal processing. The company believes that True Software Radio^{TM}will bring into being multi-mode radios that can handle multiple frequency bands, process multiple transmission protocols, be reconfigured on the fly, and be easily and cost-effectively upgraded. A fundamental step in SDR/SR technology is to get it into consumer handsets. The main problems of applying SDR/SR to cell phones are power dissipation, silicon area and price. BitWave Semiconductor Inc. claims to have a Softransceiver RFIC in 0.13um CMOS process, which uses SDR technology to enable users of cell phones, laptops and other mobile devices to communicate across different networks [5], [6]. BitWave single-chip radio frequency IC uses a traditional transceiver architecture design and adapts it to SDR-like configurability [7]. By using this approach, the system doesn't require the extremely high performance ADC and DSP of SR to provide the same flexibility under software control. Many other intermediate solutions will be proposed and coexist untill SR handsets become feasible. Business Communications Review [8] estimates that SR handset products will become commercially feasible not earlier than 2010. To make it possible earlier, continued architectural and technological innovation is required to solve the specific issues as ultra high-performance low-power ADC design. The SR in handsets requires advance in ADC and DAC technology where the major issue is battery life. The ADCs and DACs that have to operate at extremely high clock rate and the supercomputers used for digital processing consume a lot of power. The major concern is to keep the overall power dissipation within an acceptable range. Our vision of SDR in cell phones is that the technology could offer performance/power efficiency through an optimal analog/digital partitioning associated with digitally assisted analog and RF components. Instead of converting the RF signal directly to digital data it is

better to convert a high-IF signal that could provide enough flexibility, at lower sampling frequency and power dissipation.

4.   Current and Future Radio Architectures

Conceptually, the architecture of a radio receiver consists of an antenna, an RF/IF front-end, an ADC and a DSP, as shown in Figure 6.1. By placing the ADC close to the antenna, functions as filtering and frequency translation are performed in digital domain, which reduces the complexity of the receiver. However, as the ADC moves closer to the antenna the required performance of the ADC becomes very difficult to achieve [9], [10].
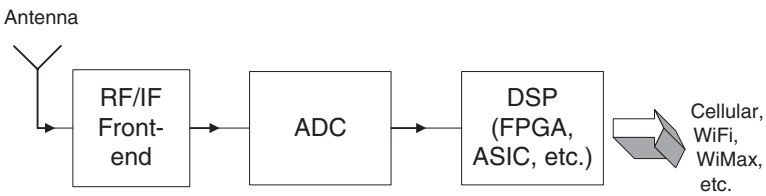
Figure 6.1.   Conceptual Block Diagram of a Radio Receiver.

The location of the ADC in a receiver chain is very important as it affects the overall performance, complexity, power dissipation, size and cost. Depending on the receiver architecture, the ADC has to digitize an RF, IF or baseband signal as is shown in Table 6.1.
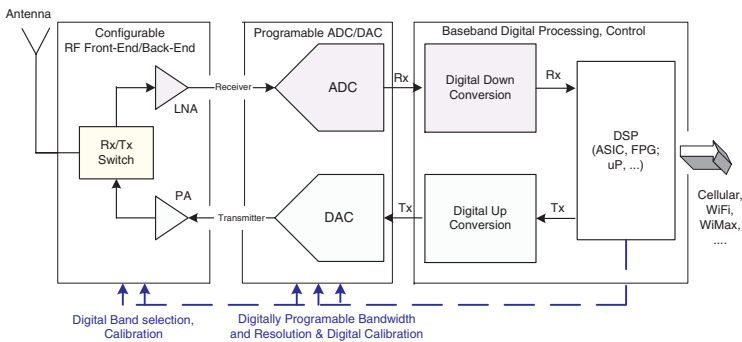
Figure 6.2.   Software Radio Architecture.

Table 6.1.   Current and Future Radio Architectures.

Radio chipset	Receiver architecture	ADC's input signal	Analog/Digital partition
Traditional Radio	Superheterodyne	I/Q Baseband IF	Dominated by the hardware and very little flexibility is provided
Reconfigurable Radio	Zero-IF/ Low-IF (near Zero-IF)	I/Q Baseband	Low IF Still dominated by hardware and only a little flexibility is provided
Software Defined Radio	High-IF	High-IF	Small amount of hardware and higher degree of reconfigurability/ flexibility is provided by software
Software Radio	Direct-RF digitization	RF	A very small amount of hardware, reconfigurability fully supported by software

4.1   Software Radio

The ultimate radio architecture is shown in Figure 6.2. The block diagram presents a generic software radio receiver and transmitter. The SR is dominated and driven by software.

In this typical SR architecture, an ultra wideband ADC is located just after the antenna, which converts RF signals directly to digital data. The down-conversion and demodulation are therefore implemented completely in digital domain on a DSP or another general-purpose processor. The SR provides full dynamic reconfigurability and programmability, but is power hungry. There-

fore, before such an architecture is reached, intermediate architectures should be considered.

4.2   Traditional Radios

The most traditional radio receivers use superheterodyne architectures. The simplified block diagram of a superheterodyne receiver is shown in Figure 6.3.
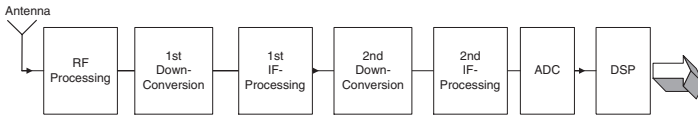
Figure 6.3.   Superheterodyne Receiver Architecture.

The requirements of an ADC in a superheterodyne architecture regarding linearity, dynamic range and bandwidth are relaxed because of the filtering and channel selection that precedes the ADC. The superheterodyne receivers are dominated by fixed hardware components and provide only a very small amount of configurability.

4.3   Alternative Approaches for SDR

Since the software radio receiver requires more mature ADC technologies, especially for mobile devices, alternative approaches to implement multi-standard receivers have to be considered. A major design issue is where to place the ADC to provide efficient radio receiver architectures. Mature alternative radio receivers such as Zero-IF and Low-IF (near zero) architectures used in commercial radio products are shown in Figures 6.4 and 6.5. They are dominated by hardware components and provide a small amount of configurability.
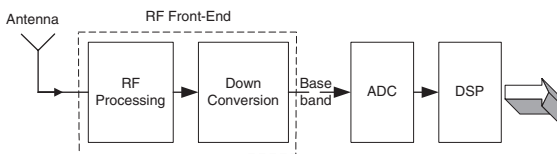
Figure 6.4.   Zero-IF Receiver Architecture.

The commercial ADC solutions for zero-IF/Low-IF receivers make use of pipelined ADC with digital calibration, but a lot of research is also ongoing
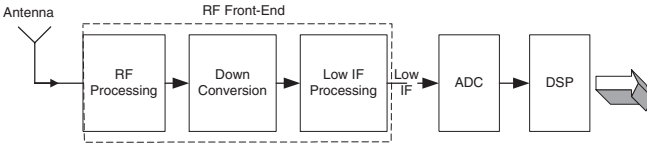
Figure 6.5.   Low-IF (near zero) Receiver Architecture.

for developing ADC architectures based on sigma-delta ($\Sigma\Delta$) modulation and other A/D conversion techniques as well. Our approaches for SDR handsets are sigma-delta ADC based because of its inherent trade-off between resolution and bandwidth, robustness to circuit imperfections, low-power dissipation, increased programmability in digital domain and higher IP re-usability. Since the ideal SDR requires an ultra high-performance ADC, a convenient alternative approach is an intermediate frequency (non-zero IF) receiver architecture where the ADC has to digitize a high-IF signal instead of a RF signal. Digitizing the frequency bandwidth at a relatively high-IF eliminates several analog stages, provides a higher degree of flexibility and is not so power hungry as in the case of a SR. Figure 6.6 shows the block diagram of such a high-IF receiver architecture. The ADC is shifted to a high-IF, which means that more analog functions are performed digitally by the DSP than in a traditional approach. The required ADC is insensitive to DC offset and low-frequency noise and the filtering in front alleviates the dynamic range, bandwidth and linearity requirements for the ADC. In addition, the high-IF receiver uses a single path until ADC and generates the I and Q paths in the digital domain to avoid I-Q analog mismatch. The issues related to the higher sampling frequencies make the high-IF ADC much less performance/power efficient, compared to a baseband ADC, but much more efficient compared to the RF digitization. Therefore, the high-IF radio architectures are well suited for today's software defined radio technology. A promising solution to enhance the overall performance/power efficiency of the wireless radio is adaptive digital compensation [11].
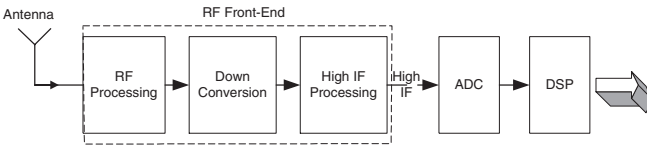
Figure 6.6.   High-IF (could be near RF) Receiver Architecture.

5.   Analog-To-Digital Conversion Challenges

In this section, the emerging and promising A/D conversion technologies that could support the SDR/SR requirements in the near and longer term are identified. The performance requirements of ADCs are pointed out, the trends in ADC technology reported recently are examined and our approaches are presented. Different enabling technologies are suitable candidates for a SR solution: multi-band antennas, multi-band multi-mode RF front-ends, multi-standard ADCs/DACs, and reconfigurable digital signal processors. One of the main challenges of the SR is the RF front-end, but the SR approach places extreme demands on the ADC as well. As it has been shown in the previous section, the receiver architecture determines how close to the antenna, the A/D conversion has to be performed. One approach to alleviate the limitation of ADC's performance/power efficiency in handsets is to use the subsampling technique [12]. In subsampling receiver architectures, the received signal is sampled with a frequency less than the IF frequency, but at least twice the data rate. While the subsampling technique eases the clock frequency requirements, it requires additional anti-aliasing filtering and increases the circuit complexity. The proper high-IF signal for achieving the highest performance/power efficiency is based on an extensive analog-digital partitioning analysis [13], [14]. A fully reconfigurable radio should cope with the existing communication standards and be able to integrate new standards. The reconfigurability requirements for the digital signal processing depend on the radio receiver architecture and selected communication standards. Suitable technologies for implementing the reconfigurable digital processing are based on: software controlled devices, like DSP; dedicated hardware, like ASIC (Applications Specific Integrated Circuit); reconfigurable hardware, like FPGA (Field Programmable Gate Arrays) and reconfigurable computing machines [9].

5.1   ADC Requirements for SDR/SR

The most important specifications for ADCs embedded in radio receivers are the sampling frequency, dynamic range, power dissipation, and linearity. A Figure-Of-Merit (FOM), based on these parameters, is used for an objective performance comparison between different ADCs. In literature, different FOMs expressing the power dissipation (Pd) of an ADC in relation to the dynamic range (DR) in a specific signal bandwidth (BW) can be found. In [14], the FOM for a Software Defined Radio's ADC has been defined as:

$$FOM = DR_{dB} + 10 \cdot \log_{10} \left(\frac{BW}{P_d} \right) \quad (dB) \qquad (6.1)$$

Table 6.2. presents ADCs with impressive performance; most of these products are not suitable for SDR/SR handsets because they are expensive and power hungry.

Table 6.2. ADCs with Outstanding Performance.

ADC Architecture	Vendor	Resolution/Bandwidth/Power
Flash ADC	Zarlink VP1058	8bit/60MHz/670mW
Sub-ranging ADC	TelASIC TC1411	14bit/75MHz/1.9W
Pipeline ADC	Linear Technology LTC2208	16bit/700MHz/1250mW
	Maxim MAX1190	10bit/400MHz/492mW
	Texas Instruments ADS5500	14bit/750MHz/780mW
	Analog Devices AD9446	16bit/540MHz/2.3W
Time-interleaved ADC	Analog Devices AD12500	12bit/70MHz/unspecified
Folding/Interpolating ADC	National Semiconductor ADC081000	8bit/1.6GHz/1.43W
Sigma-Delta ADC	Analog Devices AD7760	24bit/2.4MHz/0.958mW

Figure 6.7 shows FOM and resolution for the ADCs presented in Table 6.2. The major issue for achieving feasible FOM over wide signal bandwidth is the power dissipation. As it is illustrated in Figure 6.7, the highest resolution (24 bits) and FOM (208dB) are achieved by the sigma-delta ADC, AD7760 over the narrowest signal bandwidth (2.4MHz). By using its inherent programmability it is possible to move from the high resolution, narrow band space to the lower resolution, wider band space, while keeping the FOM in a reasonable range. The folding/interpolating ADC, ADC081000 provides a FOM of only 138dB and a resolution of 8 bits over an ultra wideband of 1.6GHz. By applying special power reduction techniques, the pipelined ADC architectures can achieve higher FOM over wide signal bandwidth.

The choice of the ADC architecture is mainly driven by:

- Radio receiver architecture and selected standards. Power dissipation, size, design cycle time and cost are the critical issues in handsets.

- Enabling capabilities of the VLSI process technology used

- Degree of flexibility, programmability and adaptability for reconfiguration.
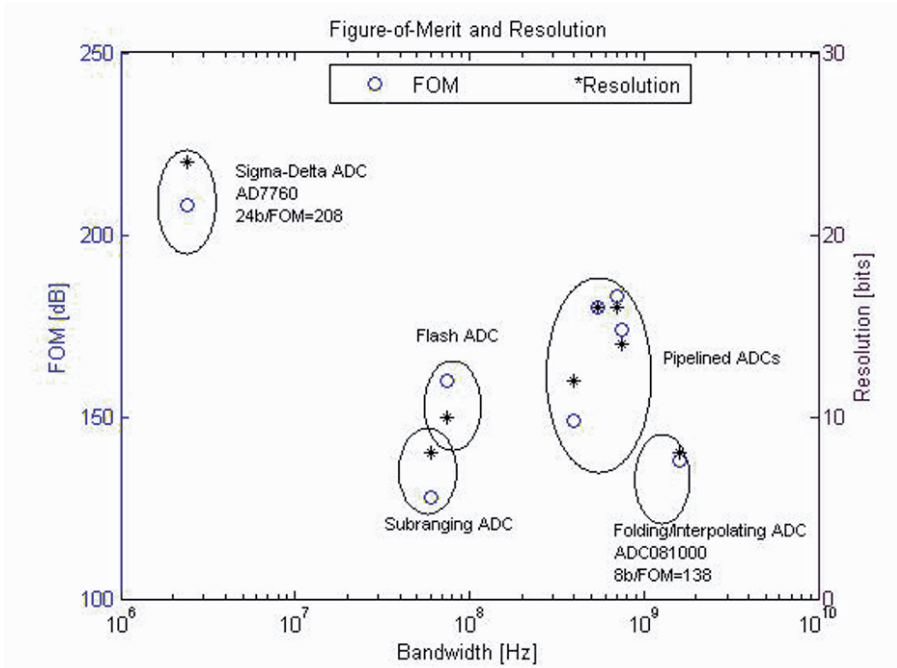
Figure 6.7.   Figure-Of-Merit and Resolution vs. Bandwidth.

The most popular wireless standards with respect to their ADC resolution requirements are presented in Figure 6.8. The performance targets for the single-standard solutions are highly optimized for low power dissipation. The major challenge for multi-standard radio receivers is the design of a reconfigurable ADC that can be configured for multiple standards while keeping the same FOM as in the case of multiple ADCs solution.

5.2   Emerging ADC Approaches

The critical issue is that the Software Radio technology requires extremely high-performance analog-to-digital converters to directly convert the signal at the antenna. The demodulation is then performed by a digital processor, which can be reprogrammed easily when standards change. Unfortunately, a full-featured software radio calls for ADCs with 16-bit or better precision operating at speeds in excess of 1 GHz. Realistic estimates suggest that even for modest dynamic range requirements, the required ADC could consume power on the order of 10 watts. Furthermore, the digital processing requires a large amount of instructions that is beyond the current DSP capabilities. As a result, without
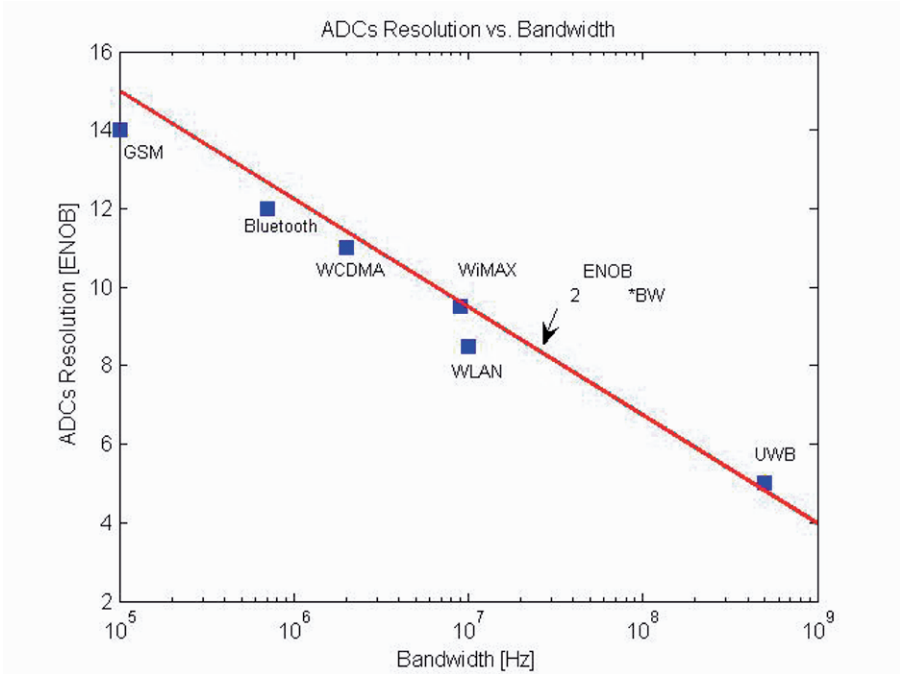
Figure 6.8. Required ADC Resolution vs. Bandwidth.

architectural and technological innovations to reduce the large and growing performance gap between analog and digital circuits, analog circuit capabilities will be the bottleneck in SR handsets development. Analog circuits can take advantage of the high performance scaled digital circuits by delegating critical design constraints to a DSP. Adaptive digital compensation techniques can be used to enhance the RF front-end and ADC performance and to relax the analog circuits' requirements [11], [15]. The enabling technologies and ADC architectures that are relevant for SDR/SR are listed in Table 6.3.

The need of ultra high performance ADCs drives the research and investigation of new technologies and architectures. Optical and superconducting ADCs can provide higher sampling frequency than silicon based ADCs, but the technologies that are feasible for implementing SDR/SR in mobile devices are those that enable the integration of the RF transceiver and the digital baseband processor on the same chip, leading to lower power dissipation, reduced die size and lower cost. Microwave technologies such as GaAs and SiGe are suitable for SDR/SR applications, but are too power hungry for SDR/SR handsets. Therefore, it is evident that CMOS technology is the most suitable one and will dominate the evolution of SDR/SR. Many modern wireless receivers exploit

Table 6.3.   Enabling Technologies and ADC Architectures for SDR/SR.

Technology	Optical	
	Superconducting	
	GaAs (gallium-arsenide)	
	SiGe (silicon-germanium)	
	CMOS	
ADC Architecture	Nyquist	Flash ADC
		Folding/Interpolating ADC
		Time-interleaved ADC
		Pipelined ADC
	Oversampling	Switched-Capacitor (SC) $\Sigma\Delta$ ADC: low-pass, bandpass
		Continuous-Time (CT) $\Sigma\Delta$ ADC: low-pass, bandpass
	ADCs Array	Flash ADC & Pipelined ADC & Sigma-Delta ADC & etc.

the inherent programmability feature of $\Sigma\Delta$ modulators [16–19], but reconfigurable pipelined ADCs [20], hybrid sigma-delta-pipelined ADC [21–23] and other ADC architectures are considered and investigated as well.

6.   Reconfigurable ADCs for SDR/SR

Our approaches for reconfigurable ADCs employ $\Sigma\Delta$ modulation. $\Sigma\Delta$ ADCs offer an inherent resolution-bandwidth trade-off, great features for flexibility, adaptability and programmability, high re-usability and integration capability. Since the linearity is a major issue in the radio receiver design, architectures with improved linearity have been proposed and designed. The theoretical dynamic range (DR) has been used in conjunction with the implementation considerations to find the optimal performance/power ratio for multi-mode operation. The theoretical DR is given by:

$$DR = f(L, N, OSR) = 10 \cdot \log\left(\frac{3}{2} \cdot \frac{2L+1}{\pi^{2L}} \cdot (2^N - 1)^2 \cdot OSR^{2L+1}\right)$$
(6.2)

where L is the modulator loop order, N is the quantizer resolution and OSR is the oversampling ratio, OSR=fS/fB , fS= sampling frequency, fB=signal bandwidth. The majority of reported $\Sigma\Delta$ADCs are implemented in CMOS technology by using SC technique. The SC circuits are very accurate since the loop filter coefficients are set by the capacitor ratios, but the sampling frequency

is limited to one-half or less of the unity-gain bandwidth of the operational amplifiers. Therefore, it will be very challenging to implement a SC $\Sigma\Delta$ ADC for future SRs. A continuous-time (CT) implementation could prove to be a practical alternative [24].

Three solutions, suitable for different radio receiver architectures are discussed in the next three subsections. They are:

- A reconfigurable switched-capacitor modified cascaded $\Sigma\Delta$ ADC suitable for Zero-IF/Low-IF receivers.

- A digitally tuned continuous-time bandpass $\Sigma\Delta$ ADC for high-IF receivers.

- A fully reconfigurable ADC for future Software Radios. This architecture is based on a field-programmable array of CT $\Sigma\Delta$ and pipelined ADCs.

6.1   A SC Sigma-Delta ADC for Traditional SDRs

A generic reconfigurable ADC based on sigma-delta modulation technology is presented in Figure 6.9. It consists of a cascaded sigma-delta ADC and a programmable DSP. The programmable DSP employs a software controller to provide different modulator architectures, sampling frequencies, resolutions, etc and to digitally enhance the ADC performance. The programmable $\Sigma\Delta$ modulator could be configured in a couple of architectures that provide the optimal performance for different wireless standards. For some wireless standards the unused blocks are switched off resulting in large power savings. Since $\Sigma\Delta$ modulators pose a high degree of programmability (loop order, quantizer resolution, oversampling ratio) a variety of architectures could be implemented based on this generic block diagram. However, a carefully selected architecture is imperative for performance/power efficiency and low cost.
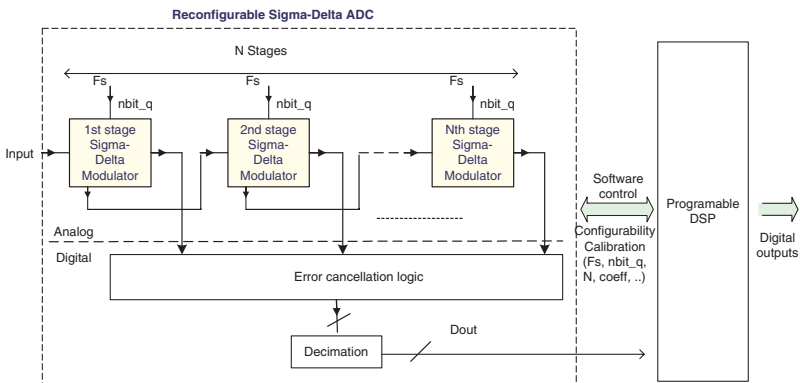
Figure 6.9.   Generic Reconfigurable $\Sigma\Delta$ ADC.
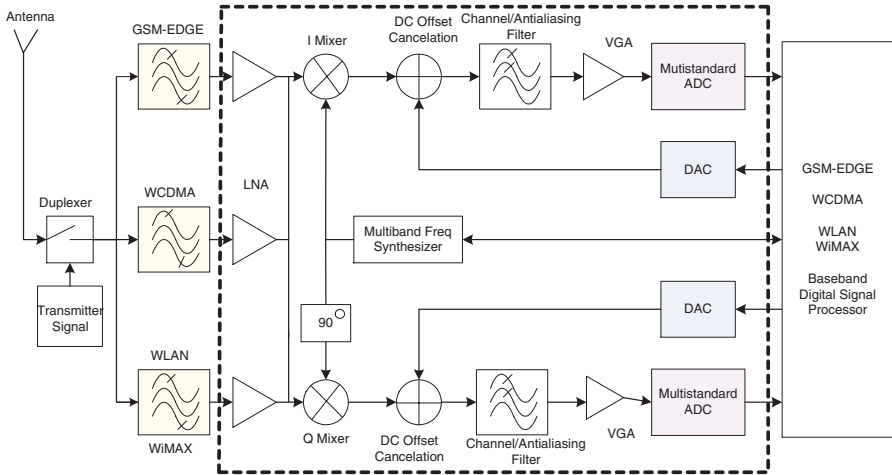
Figure 6.10.   Multi-standard Receiver.

Figure 6.11 presents our approach for reconfigurable $\Sigma\Delta$ ADC, which is optimized for the Zero-IF receiver presented in Figure 6.10. The reconfigurable ADC shown in Figure 6.11 has been designed for GSM/WCDMA/WLAN/ WiMAX standards with minimum adjustment of parameters when switching from one standard to another. The proposed modulator architecture consists of a SC 4th order 2-2 modified cascaded $\Sigma\Delta$ modulator [18]. Each stage is a 2nd order feedforward $\Sigma\Delta$ modulator that improves the linearity even at low oversampling ratio, keeping the power dissipation at an acceptable level. Decimation, error-correction and channel selection filtering are the main functions of DSP [25]. The DSP implements these functions and provides configurability for the selected standard requirements (sampling frequency, number of bits, number of stages, etc.). The first stage output, when only one bit quantizer and 1bit DAC are used, represents the GSM mode output. The output of the modified cascaded $\Sigma\Delta$ ADC represents the WCDMA mode output (for the combination 1bit-4bit) or WLAN/WiMAX modes output (for the combination 4bit-4bit). The Pseudo-Data-Weighted-Averaging (P-DWA) technique is applied in the feedback 4-bit DAC to improve its accuracy in WLAN/WiMAX modes.

The signal to noise+distortion ratio (SNDR) vs. input signal of the multistandard $\Sigma\Delta$ modulator is shown in Figure 6.12 and the performance summary is presented in Table 6.4.

The overall performance indicates that the efficiency of the proposed $\Sigma\Delta$ modulator architecture for implementing a reconfigurable ADC is quite suitable for traditional SDRs.
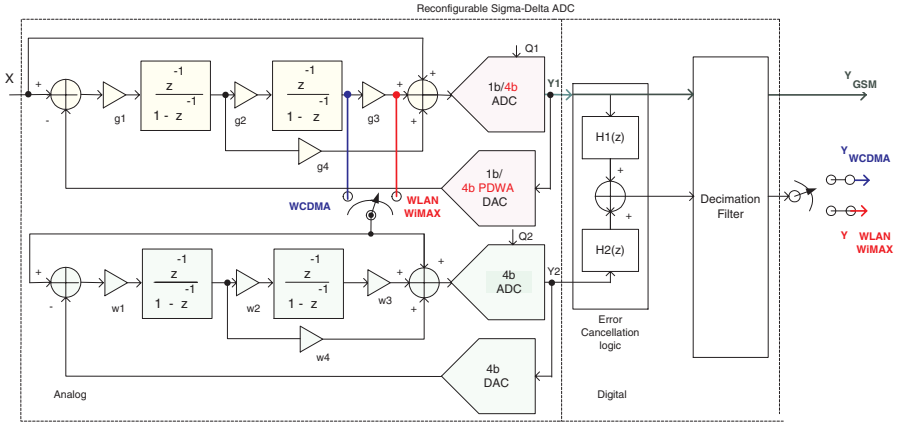
Figure 6.11.   Reconfigurable SC Modified Cascaded $\Sigma\Delta$ ADC.
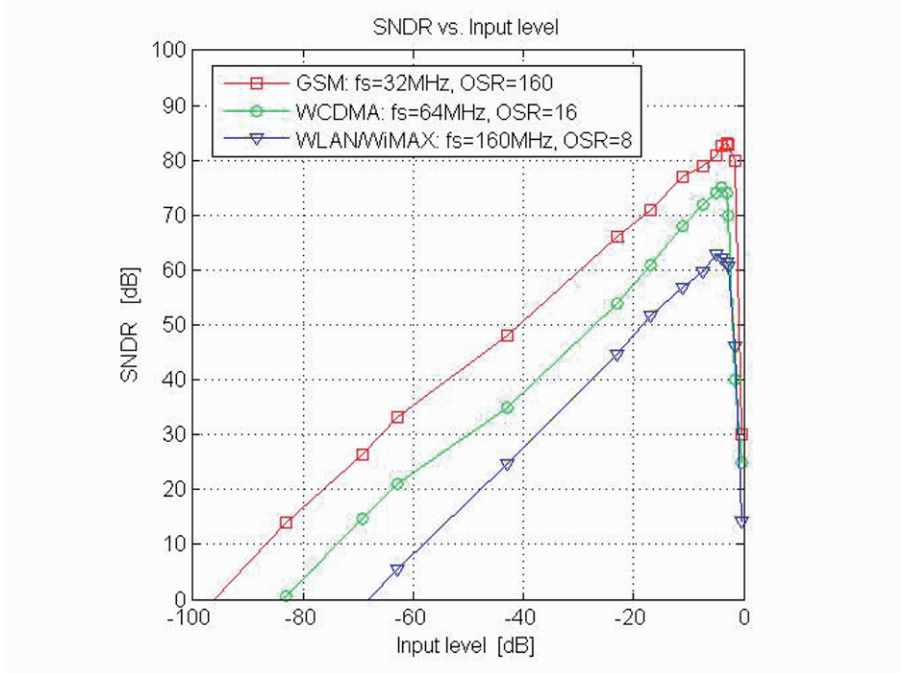
Figure 6.12.   SNDR vs. Input Signal.

Table 6.4.   Performance Summary of the Reconfigurable $\Sigma\Delta$Modulator.

Wireless Standard	GSM	WCDMA	WLAN/WiMAX
Architecture	2nd order 1b $\Sigma\Delta$ modulator with feedforward path	1b-4b (2-2) modified cascaded $\Sigma\Delta$modulator	4 b-4b (2-2) modified cascaded $\Sigma\Delta$modulator
Sampling Frequency	32MHz	64MHz	160 MHz
Bandwidth	100kHz	2MHz	10 MHz
OSR	160	16	8
Peak SNDR	83dB	75dB	62.86dB
Peak SFDR	96dB	84dB	82.2dB
IMD3	93dB	82dB	-77.5dB
Power dissipation	8.3mW	17.8mW	42mW
FOM	162	161	151
Process	0.18um CMOS, at 1.8V supply voltage		

6.2   A CT Bandpass Sigma-Delta ADC for Alternative SDRs

Sampling at the first or second IF of a receiver eliminates down conversion stages including inphase/quadrature (I/Q) mixing, but requires faster devices (higher f_T). Implementing IF processing in the digital domain takes full advantages of the increased speed, improving the noise immunity, providing more flexibility and reducing the overall power dissipation of the receiver. Additionally, performing the I/Q mixing in digital domain eliminates the I/Q channel mismatches and yields a more robust and reliable receiver. The required ADC alleviates the issues due to DC-offset and flicker noise [26]. The choice for a direct digitization at high-IF with high linearity has led to the development of a low-distortion bandpass $\Sigma\Delta$ ADC [27]. As it has been discussed earlier in this section, the SC circuits are not suitable for implementing high-IF $\Sigma\Delta$ ADCs; the sampling frequency requirement restricts the achievable IF. The CT implementation has several advantages over SC implementation, which make

them suitable for high-IF ΣΔADCs [24]. The key advantage of the CT implementation is that the sampling errors of the sample and hold (S/H) circuit are shaped as the quantization noise since the S/H is placed inside the loop. Another advantage is that a CT circuit can operate at higher frequencies, as the sampling frequency is not limited by the charge transfer accuracy requirements. Moreover, the tuning frequency of a CT filter does not depend on the sampling frequency and a CT bandpass loop filter can be tuned at a frequency other than fs/4 without requiring additional hardware. However, issues such as clock jitter and excess loop delay become great challenges to the designer, especially at high sampling frequency. Therefore, special techniques should be applied to overcome these problems. A CT bandpass (BP) ΣΔ modulator with feedforward compensation path is well suited for implementing a high-IF ADC. The block diagram for the proposed approach is shown in Figure 6.13. The high-IF ADC consists of a 4th order 4-bit CT bandpass ΣΔmodulator and a decimation filter. The decimation filtering, other digital processing, control and calibration are implemented in a DSP, which could be part of the digital baseband. By employing the bandpass ΣΔ modulator with feedforward compensation path [27] in a CT implementation, we can achieve a highly linear and low power ADC for high-IF receivers. The simulation results indicate that the proposed architecture can achieve a peak SNDR of 50dB over a bandwidth of 20MHz for an input signal centered at 75MHz and a sampling frequency of 500MHz. A digitally tuned 4th order 4-bit CT BP sigma-delta modulator can cover the dynamic range requirements for cellular, Bluetooth, WLAN and WiMAX standards.
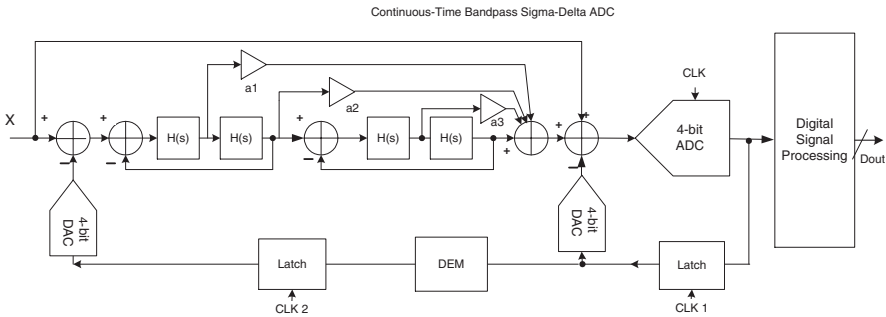
Figure 6.13. 4th order Continuous-Time Bandpass Sigma-Delta ADC.

6.3   Fully Reconfigurable ADC for Future SRs

This approach explores the possibility of implementing a fully reconfigurable ADC for future SR. The proposed approach uses the advantages of both $\Sigma\Delta$ and pipelined ADC architectures [23], [28]. Moreover, adaptive digital compensation techniques and a very powerful DSP are used to enhance the analog components performance and to make them programmable, as is shown in Figure 6.14. The DSP will choose the ADC configuration and will digitally control the analog-to-digital conversion depending on SR needs (receiver architecture, selected standard, power optimization, etc.). The DSP is also used for channel selection, demodulation and decoding. By using aggressive digitally assisted analog functions, the analog circuits could be simplified and then the power efficiency could be improved. This approach employs a programmable array of ADCs, which can cover a large range of resolution-bandwidth. The programmable array is based on CT $\Sigma\Delta$ and pipelined ADC architectures, which use the same basic building blocks (as opamps, and comparators). The ADCs array can provide a significantly large reconfigurability space and minimize the power dissipation. Similar solutions have been proposed in [21], [22].
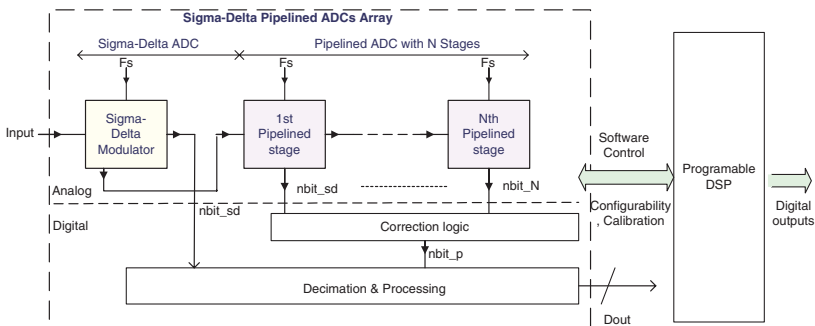
Figure 6.14.   Sigma-Delta Pipelined ADCs Array.

A dynamically configurable ADCs array can provide a wide variety of resolution-bandwidth combinations to cope with the existing wireless standards (including GSM, WCDMA, Bluetooth, WLAN, WiMAX) and future wireless standards. By using an array of ADCs, a considerable amount of power associated with the higher resolution communication standards could be conserved. The power could be saved by choosing a lower resolution in the flash ADCs or by disabling different stages for a communication standard where a lower resolution is enough. Figure 6.15 shows the block diagram of our proposed CT $\Sigma\Delta$-pipelined ADCs array.
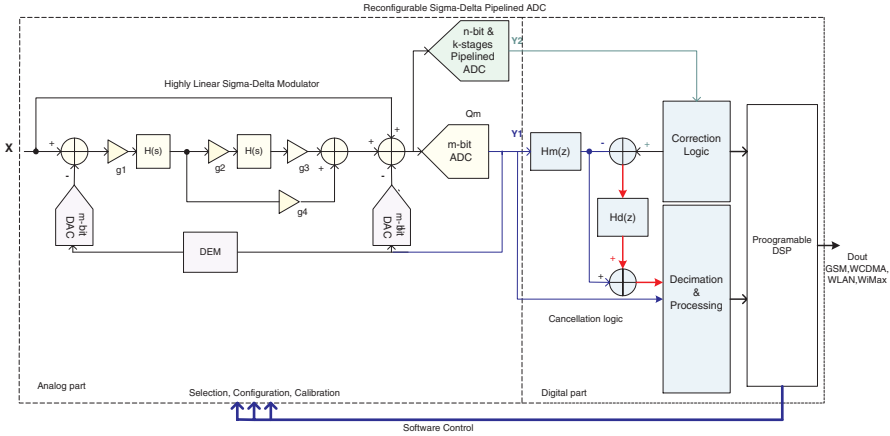
Figure 6.15. CT Sigma-Delta Pipelined ADCs array.

The ADCs array can be configured as CT $\Sigma\Delta$ADC or pipelined ADC or cascaded CT $\Sigma\Delta$-pipelined ADC depending on the communication standard requirements. It is also possible to change the sampling frequency according to the input signals that have to be process in a specific wireless standard. A software controller block takes the decision of changing the resolution or sampling frequency, disabling different stages or blocks according to a cellular, WLAN or WiMAX communication standard. Obviously having a ADCs array with a large space of reconfigurability and high degree of adaptability it will be possible to upgrade and add new standards over time.

7. Conclusions

Next generation wireless systems will have to support a wide range of data rates over several signal bandwidths and will require flexible system resources. The ADC is the key component towards a fully integrated software radio. This chapter has provided an overview of the emerging analog-to-digital conversion technologies for SDR/SR handsets. The SDR concept was introduced and the enabling radio architectures were presented. The ADC design challenges for SDR/SR handsets were identified and finally, our reconfigurable ADC approaches for traditional SDR, alternative SDR and future SR handsets were presented. This chapter has shown that a Software Defined Radio implementation for mobile terminals is within reach. The major remaining challenges are the degree of configurability, programmability and adaptability and the further reduction in power dissipation, size and cost. Many additional challenges come when SDRs/SRs design moves to nanometer technologies. As we enter into the nanoscale era, the key question is whether the capability of new nanometer

processes, such as 65nm CMOS or smaller would allow robust, manufacturable solutions that maintain signal integrity over random process and environmental variations.

Acknowledgments

References

[1] John Mitola III. Software Radios-survey, critical evaluation and future directions. *IEEE Aerospace and Electronic Systems Magazine*, 8(4):25–36, April 1993.

[2] Vanu Inc. http://vanu.com/technology/softwareradio.html.

[3] Ronald M. Hickling. New technology facilitates true software-defined radio. *RF Design*, pages 18–26, April 2005.

[4] TechnoConcepts. http://www.technoconcepts.com.

[5] BitWave. http://www.bitwavesemiconductor.com/technology.htm.

[6] J. A. Kilpatrick and et al. New SDR architecture enables ubiguitous data connectivity. *RFDesign*, pages 32–38, January 2006.

[7] Kevin Morris. Redefining Software Defined Radio- BitWave's SDR for the Masses. *Embedded Technology Journal*, 2005.

[8] Business Communications Review. http://www.bcr.com/bcrmag/2005/04/p18s1.php.

[9] D. Grandblaise and K. Moessner, editors. *Requirements on Reconfigurability for Dynamic Spectrum Allocation*. 2004.

[10] E2 R White Paper. Hardware Technology Exploration: Impact of Technology Evolution on End to End Reconfigurability. *http://e2r.motlabs.com/whitepapers*, December 2005.

[11] A. Reza Rofougaran, M. Rofougaran, and A. Behzad. Radios Next-Generation Wireless Networks. *IEEE Microwave Magazine*, pages 38–43, March 2005.

[12] M.A.L. Mostafa, M.C. Fernando, W.K. Chan, and C. Gore. WCDMA receiver architecture with unique frequency plan. In *IEEE ASIC/SOC Conference*, pages 57–61, 2001.

[13] P. B. Kenington and L. Astier. Power Consumption of A/D Converters for Software Radio Applications. *IEEE Trans. on Vehicular Technology*, 49(2):643–650, March 2000.

[14] R. Schreier. ADCs and DACs: Marching Towards the Antenna. *IEEE ISSCC - Girafe Workshop*, pages 1–19, February 1003.

[15] B. Murmann and B. Boser. Digitally Assisted Analog Integrated Circuits. *Closing the gap between analog and digital - Focus DSPs*, pages 65–71, March 2004. www.acmqueue.com.

[16] J. Koh, K. Muhammad, B. Staszewski, G. Gomez, and B. Horoun. A Sigma-Delta ADC with a built-in Anti-aliasing filter for Bluetooth receiver in 130nm digital process. *IEEE CICC*, pages 535–538, 2004.

[17] J. Arias and et al. A 32-mW 320-MHz Continuous-Time Complex Delta-Sigma ADC for Multi-Mode Wireless-LAN Receivers. *IEEE Journal of Solid-State Circuits*, 41(2):339–351, February 2006.

[18] A. Rusu, D. Rodríguez de Llera González, and M. Ismail. Reconfigurable ADCs enable smart radios for 4G wireless connectivity . *IEEE Circuits and Devices Magazine*, May 2006.

[19] R.H.M. van Veldhoven R. A triple-mode continuous-time $\Sigma\Delta$ modulator with switched-capacitor feedback DAC for a GSM-EDGE/CDMA2000/UMTS receiver. *IEEE Journal of Solid-State Circuits*, 38(12):2069–2076, December 2003.

[20] A. Bahai. Pipelined analog-to-digital converter that is configurable based on wireless communication protocol. *US Patent 6980148*, December 2005.

[21] K. Gulati and Hae-Seung Lee. Reconfigurable analog-to-digital converter. *US Patent 6864822*, March 2005.

[22] G. Gielen. Reconfigurable A/D converters for 4G radios. Technical report, presentation at Royal Institute of Technology Stockholm, Sweden, November 2005.

[23] A. Rusu. Wide bandwidth analog-to-digital converters: Architectures and topologies based on $\Sigma\Delta$ modulation and pipeline algorithm. Technical report, unpublished work presented at Royal Institute of Technology Stockholm, Sweden, December 2001.

[24] Yann Le Guillou. Analyzing sigma-delta ADCs in deep-submicron CMOS technologies. *RF Design Magazine*, pages 18–26, February 2005.

[25] P. Kiss, J. Silva, A. Wiesbauer, T. Sun, U-K Moon, J. Stonick, and G. Temes. Adaptive digital correction of analog errors in MASH ADCs - Part II. Correction using test-signal injection. *IEEE Transactions on Circuits and Systems, part II*, 47:629–638, July 2000.

[26] L. Breems and J. H. Hijsing. *Continuous-time sigma-delta modulation for A/D conversion in radio receivers*. Kluwer Academic Publishers, 2001.

[27] A. Rusu and M. Ismail. A Low-Distortion Bandpass $\Sigma\Delta$ Modulator for Wireless Radio Receivers. *IEE Electronics Letters*, 41(19):1044–1046, September 2005.

[28] T.L. Brooks and et al. A cascaded sigma-delta pipeline A/D converter with 1.25 MHz signal bandwidth and 89 dB SNR. *IEEE Journal of Solid State Circuits*, 32(12):1896–1906, December 1997.

Chapter 7

RECONFIGURABLE A/D CONVERTERS FOR FLEXIBLE WIRELESS TRANSCEIVERS IN 4G RADIOS

Georges Gielen, Erwin Goris and Yi Ke

1. Introduction

Flexibility is a key feature in 4G telecom systems, where there is a demand for reconfigurable transceivers that can cope with multiple standards (cellular, WLAN, Bluetooth, etc.). Additionally, even within one mode, these transceivers should adapt to the environment (presence of received blockers or not, status of battery power levels, etc.) to minimize power consumption and optimize performance according to the needs of the customer and the desired Quality of Service. In addition, flexibility is required to cut the development time and cost to implement possible new future standards into the 4G system. All this calls for a digitally-controlled front-end architecture ("software-defined radio") with reconfigurable RF and analog baseband blocks controlled through digital programmable software. This poses serious challenges to the design of such reconfigurable yet power-efficient RF/analog blocks. For the analog-to-digital converters in the receiver, this comes down to designing a power- and area-efficient reconfigurable converter with variable bandwidth and variable dynamic range. The general requirements for such converters in 4G systems will be described in this chapter. This will then be illustrated with the design of a reconfigurable continuous-time $\Delta\Sigma$ A/D converter with a pipelined multi-bit quantizer and 1-bit feedback. The prototype chip has been realized in a 0.18 μm CMOS technology. It has 3 different modes (20 MHz BW/58dB SNDR, 4 MHz BW/60dB SNDR, 0.2MHz BW/70dB SNDR). The chip has an active area of 0.9 mm^2 and the power consumption for the most demanding mode (20 MHz/58 dB) is 37 mW.

123

2.   Towards 4G Radios

Few technological achievements have had a more spectacular worldwide impact than the digital cellular phone, which has reached an enormous market penetration today. In addition, many other wireless applications are becoming true success stories. WLAN (IEEE 802.11x) is gaining wide popularity in urban areas, companies and homes, with many WiFi hot spots being installed for public and commercial use. Many cars are being equipped with GPS receivers, and Bluetooth is used for short-range communication such as between your cell phone and a fixed transceiver on the dashboard of your car. Radio and TV broadcasting will become digital in the near future (DAB, DVB-T/H). 3G cellular communication systems are currently being deployed, and future 4G systems are under development. Obviously, mobility and wireless connectivity seem to satisfy some basic human needs.

It is clear from this discussion that there is a large and still growing number of wireless standards for all kinds of applications (long-range, short-range, voice, data, images, broadcast, on-line games, etc.). Table 7.1 gives an overview of different wireless standards and some basic requirements they pose on the baseband analog-to-digital converters. To cope with all these standards, customers don't want a collection of different handheld devices, but they want all services to be integrated into one portable device which can offer mobility (through roaming and seamless hand-over between services, e.g. between WLAN inside and cellular when leaving the office building) and which can simultaneously handle different services (e.g. operate GSM and GPS at the same time). The future 4G systems promise to offer customers an integration of such different services. So convergence of services is a very important aspect of 4G, besides the mere promise of higher data rates (100+ Mbps).

A second feature of 4G systems is the scalable optimization of power consumption. The power consumption of these transceivers has to be minimized

Table 7.1.   Different wire standards and some basic ADC requirements.

Standard	Bandwidth	Resolution
GPS	2 MHz	10 bit
GSM	200 kHz	12 - 14 bit
Bluetooth	1 MHz	13 bit
WCDMA	3.84 MHz	6 - 8 bit
WLAN 802.11a	20 MHz	10 - 14 bit
WLAN 802.11b	22 MHz	6 - 8 bit
WLAN 802.11g	22 MHz	10 - 14 bit

and adapted to the environment (presence of received blockers or not, status of battery power levels, etc.) in order to optimize performance at minimum power according to the needs of the customer and the desired Quality of Service. For example, when the battery levels are running low, the transceiver may switch to a lower operating mode to keep the transmission running, be it at lower quality. Similarly, when for instance no blocking signals are present, the system can switch to a mode of operation with lower dynamic range, hence saving power, while switching back to a mode with higher dynamic range when blockers are detected in the received antenna signal. The transceiver therefore needs to be dynamically adaptive, driven by the quality of service requested by the user, and not only switch configuration in a static sense when switching communication standard.

   Instead of having a separate receiver and handheld device for every standard, customers want to have all functionality integrated in one multi-standard device. In order to build such a multi-standard adaptive transceiver, different solutions exist. The most straightforward way (see Fig. 7.1(a)) is to integrate several dedicated transceivers in parallel, one for each standard. This may be power efficient since every standard has an optimized transceiver (while the others are turned off), but it is a very expensive solution in terms of chip area and cost. Another approach is to use one global transceiver, where the building blocks are designed for the worst-case set of specifications. This is very inefficient, since it will consume way too much power for most of the standards and operation modes, and for a broad range of specifications this solution is simply technically infeasible. Therefore, a transceiver with reconfigurable building blocks (see Fig. 7.1(b)) where the performance and hence the power consumption can be scaled to the needs of the standard and operating mode, is much more

(a)   (b)
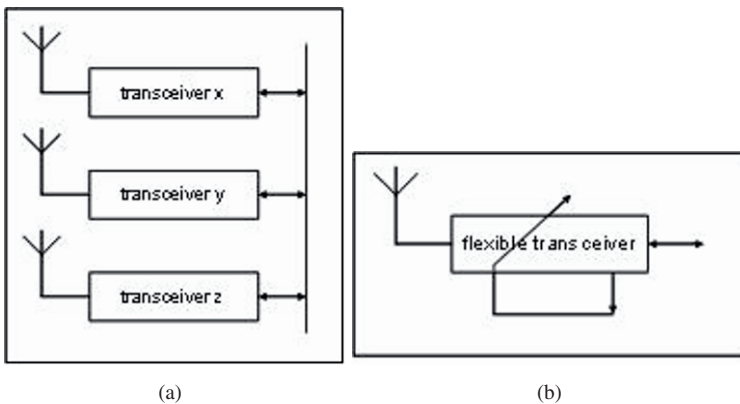
Figure 7.1. (a) Multiple parallel transceivers for each standard, versus (b) one flexible transceiver for multiple standards.

desirable. This is the concept of a "flexible" or "reconfigurable" transceiver, which has gained considerable interest recently. If such reconfigurable front-end is controlled and reconfigured through digital programmable software, then this is called a "software-defined radio".

Section 2 describes the concept of flexible receivers. Section 3 focuses on the A/D converter and its specifications in such a flexible 4G receiver. Existing designs are compared and a new topology is proposed. Section 4 describes the design of a prototype chip in a 0.18 μm CMOS technology. Conclusions are drawn in section 5.

3.   Flexible Receiver Architectures

An ideal flexible receiver should be able to cope with a wide range of communication standards. It should adapt to the environment to optimize power consumption and performance according to the needs and quality of service requests of the customer. Flexibility should also cut the development time and cost to implement and integrate future new standards, since such new standard can reuse, possibly after some reconfiguration, the existing front-end blocks (provided that the reconfigurable blocks can manage the requested performance range). It should be stressed that a flexible receiver is not a "software radio". Software radio implies that almost all analog signal processing is shifted to the digital domain, and that the front-end consists of a large A/D converter right after the antenna. This puts extreme requirements on the A/D converter (both bandwidth and dynamic range) and leads, if even feasible, to a very power-inefficient solution for most operations.

A flexible receiver consists of a digitally controlled analog front-end ("software-defined radio") and a programmable digital back-end. As an example, a simplified schematic of a zero-IF flexible receiver is shown in Fig. 7.2. The digital back-end processes the signals and feeds back control signals that can reconfigure (statically between different standards or even dynamically within the same standard) the building blocks in the front-end. These blocks then switch to a different set of performance values (e.g. a different resolution and bandwidth for an analog-to-digital converter), or a different filter order or cut-off frequency for a filter, or a different gain and bandwidth for a low-noise amplifier, or a different gain for a variable gain amplifier, etc.). This poses significant challenges on the design of reconfigurable RF and analog baseband blocks, which should have close to minimal power consumption in all performance combinations compared to dedicated implementations of the same block for the same set of performance values. Note that only one antenna is drawn in Fig. 7.2, but more likely an array of antennas (MIMO) will be used. The same remark applies to the band-select filters, which are generally implemented as off-chip SAW filters. Note also that in reality in a 4G system still more than one transceiver in parallel might be needed for communication services that
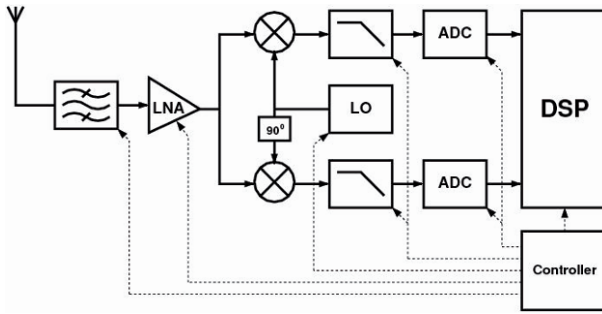
Figure 7.2. Simplified block diagram of a flexible software-defined receiver with digital control of a reconfigurable front-end.

need to be able to operate simultaneously at any time (e.g. GPS reception might be needed in parallel to voice/data communication) and that cannot be time-interleaved.

The differences between existing multi-standard transceivers and a flexible 4G one are the granularity and the range of building block specifications covered. Numerous IEEE 802.11a/b/g transceiver designs have already been published and/or are commercially available. Another popular combination is GSM-EDGE/WCDMA. A flexible 4G receiver should be able to deal with the low-bandwidth/high-dynamic-range requirements of both cellular standards, as well as the high-bandwidth/low-dynamic-range requirements of WLAN standards, in combination with other standards like Bluetooth and DVB.

We will now focus on the reconfigurable A/D converters in such flexible transceivers. To cope with different standards (cellular, WLAN…), the A/D converter needs to have a variable bandwidth and dynamic range. $\Delta\Sigma$ A/D converters are quite suitable for this task. They allow an easy trade-off of bandwidth and dynamic range. They also lower the specifications for the channel/anti-aliasing filter in front of the ADC. But their use is challenging for applications with a wider bandwidth like WLAN, which rather call for pipelined type of converters.

4. Multi-standard A/D Converters

4.1 Different Architectures

The typical bandwidth and dynamic range (number of bits) requirements for different wireless communication standards were included in Table 7.1. The bandwidth ranges from between 0.2 to a few MHz on the one end up to 22 MHz on the other end. The intrinsic performance of analog circuits and hence also of analog-to-digital converters is limited by a trade-off between speed, accuracy and power resulting in a relation of the form [1]:
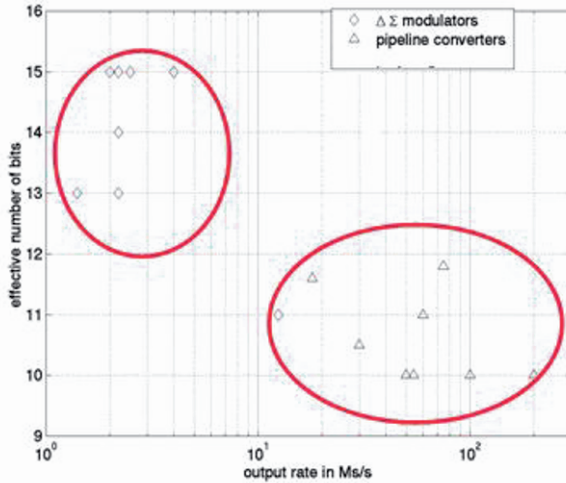
Figure 7.3.   Comparison of performance of $\Delta\Sigma$ and pipelined A/D converters.

$$\frac{\text{speed} * \text{accuracy}^2}{\text{power}} = \text{const} \tag{7.1}$$

This implies that converters can be either fast but less accurate, or more accurate but slower for the same power. Hence, if we compare the speed (sample rate) and the accuracy (effective number of bits) of several published A/D converters, as shown in Fig. 7.3, then we notice that the large-bandwidth ones are pipelined converters, while the higher-resolution ones are $\Delta\Sigma$ converters.

Therefore, for cellular standards (e.g. GSM-EDGE, WCDMA) with medium to high dynamic range requirements and a fairly limited signal bandwidth, an oversampling converter is commonly used. Powerful blocker signals can exist near the signal band. This requires a filter with very sharp edges, hence with a large power and hardware cost. One of the main advantages of an oversampling converter is that it relaxes the specifications of the analog channel filter in the receiver. Channel filtering is performed in the digital domain at a reduced cost. If a continuous-time $\Delta\Sigma$ A/D converter is used, the loop filter also acts as an anti-aliasing filter.

For WLAN standards (e.g. IEEE 802.11a,b,g) on the other hand, with medium dynamic range requirements but with a fairly high bandwidth (> 10 MHz), a Nyquist-rate converter like a pipelined ADC is commonly used. Reconfiguration of such pipelined converter can occur through switching on/off some stages (see Fig. 7.4) [2], changing the sampling rate, through reconfiguring the amplifiers (transistor sizes, bias currents, capacitors) in the stages (see Fig. 7.5), etc. The impact of the many switches on the circuit performance is a concern,
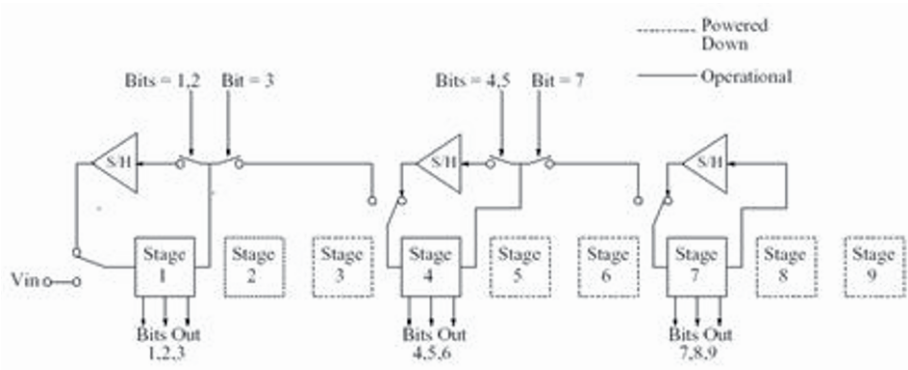
Figure 7.4. Changing the resolution and bandwidth of a pipelined A/D converter by switching in/out stages and by changing the sampling rate.
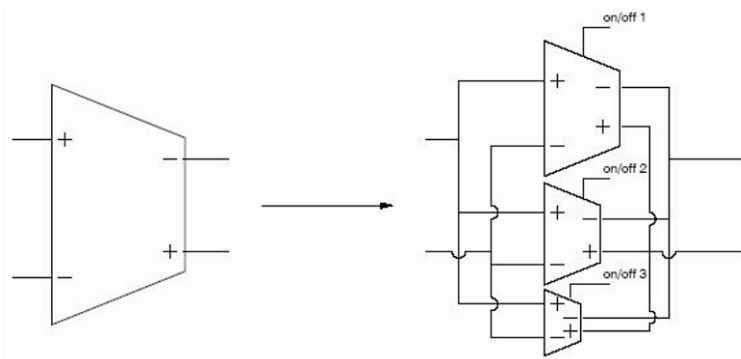
Figure 7.5. Reconfiguring an OTA by switching in/out different substages (different sizes, currents or capacitors.

however. In addition, pipelined converters are quite power-hungry blocks. In recent years, however, some $\Delta\Sigma$ A/D converter designs have been published with a signal bandwidth higher than 10 MHz. Most of these converters use a continuous-time loop filter [3] [4], but a switched-capacitor implementation has also been published [5]. Like for cellular standards, using an oversampling converter, if feasible, would lead to a more power-efficient design, and they allow an easy trade-off between bandwidth and dynamic range. In addition, they scale well with technology.

Fig. 7.6 shows a general block diagram of an oversampling $\Delta\Sigma$ A/D modulator consisting of a loop filter H(s), a quantizer and a DAC feedback block. The quantizer is operated at a frequency much higher than the signal's Nyquist frequency (the ratio is called the oversampling ratio OSR), which spreads the
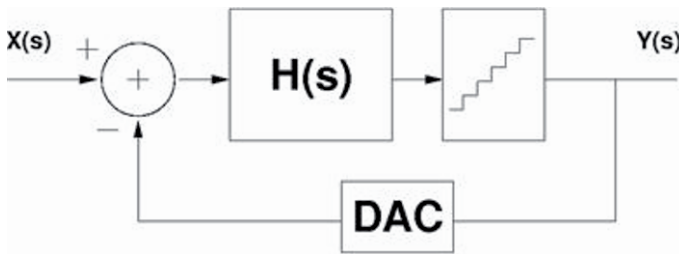
Figure 7.6.   Basic Block Diagram of a $\Delta\Sigma$ modulator.

quantization over a larger frequency range, and the loop filter shapes the quantization noise to higher frequencies, which is then removed by the digital decimation filter that follows the modulator, resulting in larger SNR values.

Most published reconfigurable A/D converters today combine 2 or 3 cellular standards (e.g. UMTS/GSM [6], WCDMA/GPRS [7], GSM-EDGE/CDMA2000/UMTS [8]). As can be expected for these combinations with a relatively low signal bandwidth, they all employ oversampling. Switching between different modes is then done by changing one or more of the following properties: loop-filter coefficients, loop-filter order, OSR, sampling frequency, number of bits in the quantizer. For converters with already a high signal bandwidth (e.g. 10 MHz), only a limited OSR is possible due to technological limits. For a continuous-time converter, increasing the OSR also means increasing the jitter sensitivity. Power and area consumption of these reconfigurable flexible converters are comparable with state-of-the-art dedicated converters. This shows that the $\Delta\Sigma$ architecture is very well suited for multi-standard A/D converters. WLAN standards, however, which have a much larger signal bandwidth, generally use a Nyquist-rate converter. In the flexible ADC topology proposed in this chapter, that is targeting the combination of cellular and WLAN standards as summarized in Table 1, oversampling is also used in the wide-bandwidth mode.

One of the few published A/D converters with a very widely programmable bandwidth and dynamic range was presented in [9]. It has a pipelined mode and a $\Delta\Sigma$ mode. The bandwidth range is 0-10 MHz. The resolution range is 6-16 bits. Opamps and capacitors are shared between pipelined and $\Delta\Sigma$ mode. Because of the large reconfigurability space and the effective combination of two architectures into one combined architecture, the design is very complex with a considerable area overhead.

4.2   Wide-bandwidth $\Delta\Sigma$ A/D Converters

The flexible oversampling ADC topology that has been developed is now discussed in more detail. It we were to apply a $\Delta\Sigma$ converter for the desired

reconfigurable multi-mode converter, then the first parameter to scale would be the oversampling ratio (OSR). To first-order approximation, increasing the OSR does not increase power consumption: the kT/C noise is spread out over a broader bandwidth, so the capacitors can be scaled down. However, if a wide signal bandwidth (say 20 MHz) is required, only a limited OSR is possible due to technological limits. To increase the resolution, basically two other parameters can be altered: the filter order and the resolution of the quantizer. Increasing the filter order from say second to third order for a single-loop converter with single-bit feedback is only helpful when the OSR is larger than 16. This can clearly be seen in Fig. 7.7, which shows the SNDR as a function of the OSR for different $\Delta\Sigma$ converter topologies ("O n, m b" denotes "order n, m bits").

Increasing the number of bits of the quantizer on the other hand greatly improves the performance of the converter. As can be seen in Fig. 6, going from a 1-bit to a 2-bit implementation improves the total SNDR with 14dB ! This is much more than the 6 bit predicted by a simple linearized $\Delta\Sigma$ A/D converter model. This is explained by the fact that a multi-bit converter allows higher overload levels than a single-bit one.

The price paid for this extra performance however is substantial. The main problem is the linearity requirement of the feedback DAC. For a single-bit implementation, this DAC is inherently linear. This is no longer the case for the multi-bit implementation. The distortion components due to the DAC non-linearity have the same transfer function to the output as the signal, so they are present in unsuppressed form in the output spectrum. Generally, Dynamic
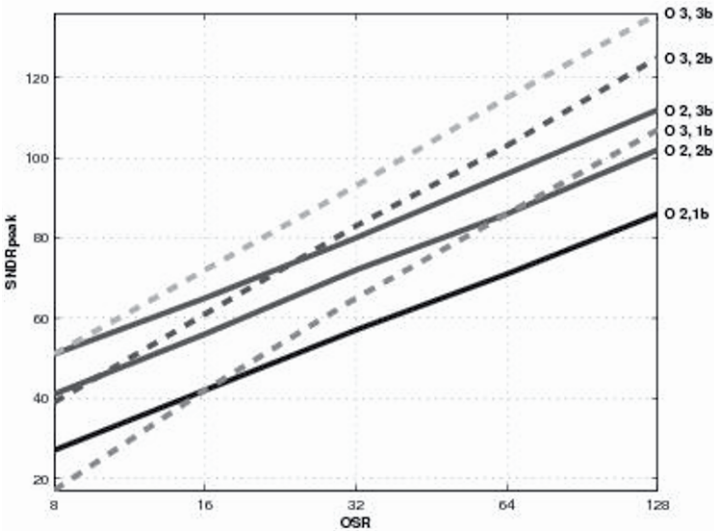
Figure 7.7. SNDR vs. OSR for different $\Delta\Sigma$ A/D converter topologies.

Element Matching (DEM) techniques are employed to whiten and shape the distortion components. These techniques can become quite complex and introduce extra delay in the feedback path. Certainly for high-speed converters they become difficult to implement.

In the Leslie-Singh architecture [10] (see Fig. 7.8) only the most significant bit of the multi-bit quantizer is fed back to the input. This has some interesting advantages. There are no DAC linearity problems and the quantizer can be pipelined. In the figure two separate quantizers are shown, but in practice the single-bit quantizer can be the first stage or the MSB of a pipelined ADC.
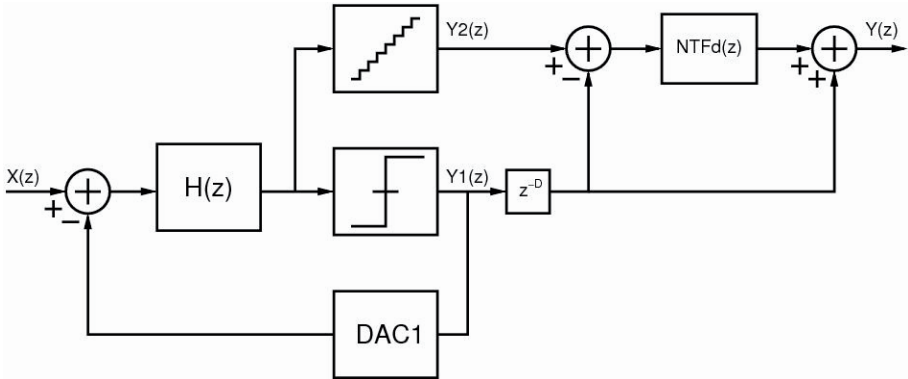
Figure 7.8.   The Leslie-Singh architecture with discrete-time loop filter and digital reconstruction filter.

There are however some disadvantages associated with this architecture. The main disadvantage is the need for a digital filter to combine both digital outputs. This digital filter should have a frequency characteristic (NTFd) identical to the (analog) noise transfer function (NTF). This can be derived as follows. The output of the modulator is given by:

$$Y(z) = z^{-D}STF(z)X(z) + NTF_d(z)Q_2(z)$$
$$+z^{-D}(NTF(z) - NTF_d(z))Q_1(z) \tag{7.2}$$

where Q1(z) is the quantization noise of the first quantizer, Q2(z) is the much smaller quantization noise of the second multi-bit quantizer, and STF(z) is the signal transfer function of the modulator. Looking at equation (2), we see that if the NTF of the modulator equals the transfer function of the digital filter NTFd, the quantization noise of the first quantizer will be removed from the output. The quantization noise at the output will then be the much smaller shaped quantization noise of the second quantizer. This architecture will not give the advantage of an increased overload level when increasing the resolution

of the second quantizer. Adding a bit in the second quantizer will add only 6 dB to the SNDR.

Matching the two filter characteristics NTF and NTFd however is not straight-forward. For a discrete-time loop filter this requires sizing the capacitors to ensure sufficient matching. This costs power and area. To relax the match-ing constraints, more than one bit can be fed back [12]. This however implies sacrificing one of the major advantages of the topology: no DAC linearity prob-lems. If a continuous-time loop filter is used, accurate matching of the filter coefficients is simply impossible.

A different approach was suggested in [13]. That paper presents a design with an adaptive digital filter. To adapt the coefficients in the filter, a test signal is injected in front of the quantizer. This is the least sensitive node in the modulator. The test signal experiences the same transfer function to the output as the quantization noise of the first integrator. The error signal that guides the adaptation process is the correlation between the test signal and the actual output. If these two signals are uncorrelated, then all the quantization noise of the first quantizer is removed from the output.

Another approach based on the same principle would be a simple off-line blind calibration. This works remarkably well for a second-order loop filter as in this case only 2 coefficients need to be adapted. The principle is shown in Fig. 7.9. A digital IIR filter (implemented in the DSP block) is used with two integrators. The filter's block diagram is shown in figure 7.10(a). The algorithm used to adapt the coefficients in our design was a very simple hill
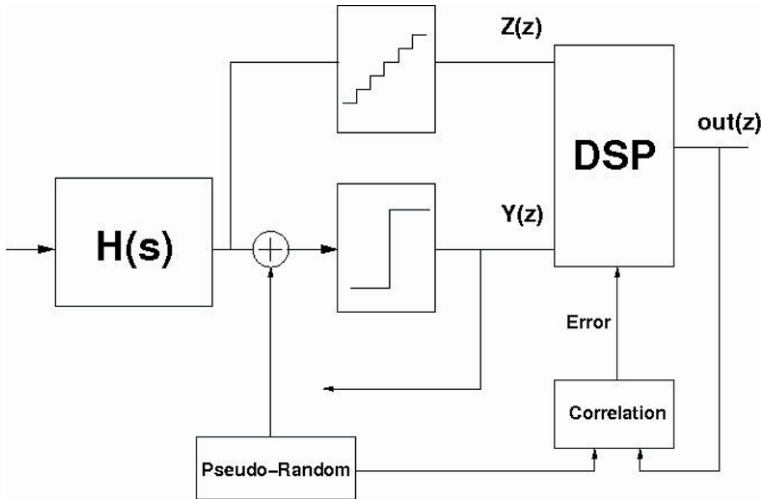
Figure 7.9. Principle schematic of the tuning of the digital filter (implemented in the DSP block).

(a)

(b)
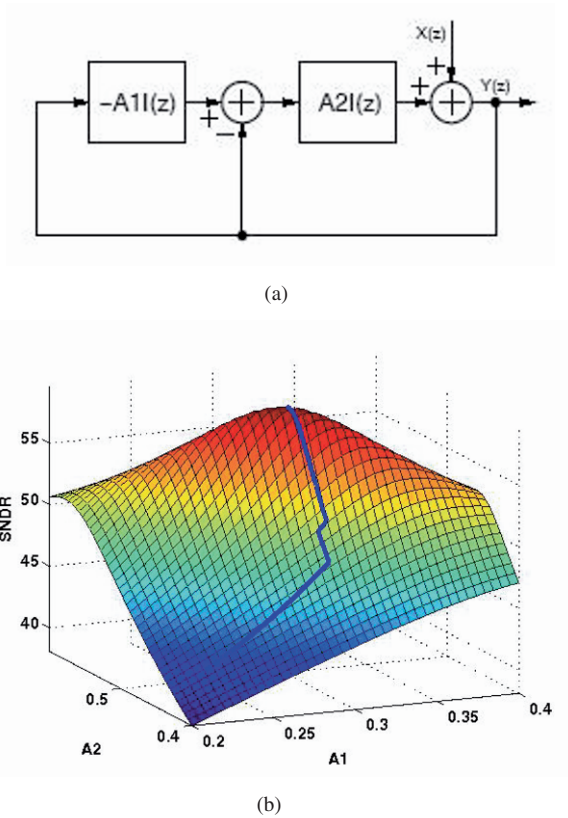
Figure 7.10. (a) Structure of the digital IIR tuning filter, and (b) impact on the SFDR by tuning the digital filter coefficients.

climbing algorithm with a fixed step size. Of course, more elaborate algorithms are also possible. The calibration is performed off-line with no signal present at the input. The coefficients are adapted in such a way that the low-frequency quantization noise is minimized. Figure 7.10(b) shows the adaptation of the coefficients and the impact it has on the SNDR. This approach was used in our prototype design.

4.3 Prototype Reconfigurable Wide-bandwidth Continuous-time $\Delta\Sigma$ A/D Converter

The presented architecture consists of a continuous-time $\Delta\Sigma$ A/D converter with a pipelined quantizer and 1-bit feedback (see Fig. 7.8). This is the Leslie-Singh topology [11], but implemented with a continuous-time loop filter. To ensure matching between the analog filter and the digital reconstruction filter,
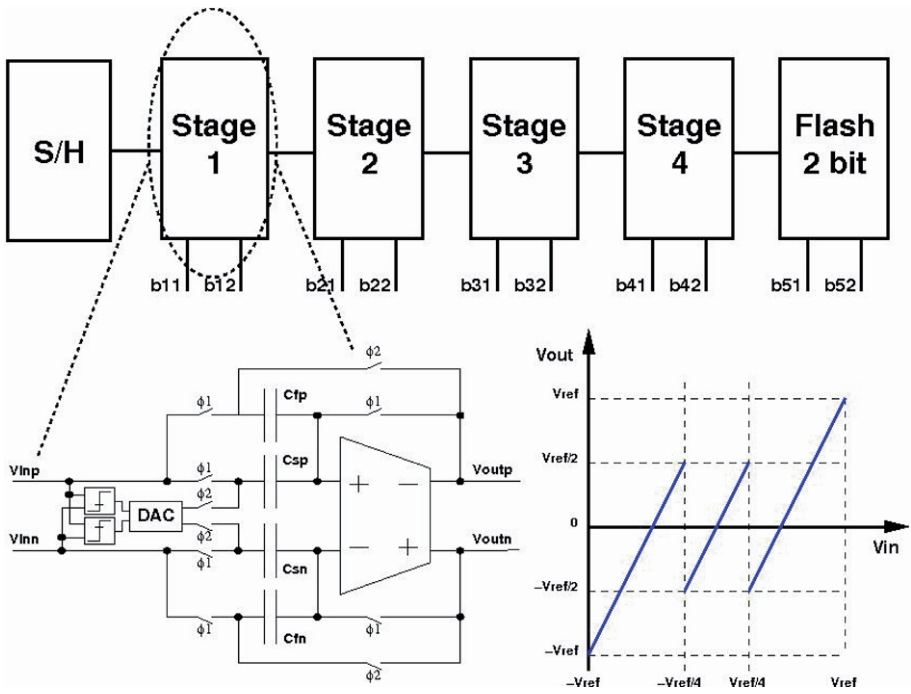
Figure 7.11. Circuit schematic of the pipelined quantizer used in the A/D converter.

a simple blind digital off-line coefficient adaptation scheme is used (see Fig. 10).

A prototype chip implementing this architecture has been designed in a 0.18 μm CMOS technology. The converter has 3 different modes (20MHz BW/58dB SNDR, 4 MHz BW/60dB SNDR, 0.2MHz/70dB SNDR). For the wide bandwidth (20 MHz), a small oversampling ratio of 12 has been used. This implies a very challenging sampling frequency of 480 MHz. To increase the dynamic range in this mode, the quantizer is a 6-bit pipelined A/D converter (see Fig. 7.11).

Every pipeline stage in the pipelined quantizer has a resolution of 1.5 bits. Only the first bit of the quantizer is fed back as explained in section 3.2. Using a pipelined A/D converter has the extra advantage that the resolution and hence the power consumption can be lowered easily by switching of pipeline stages at the end of the converter. An amplifier bandwidth of 2.5 GHz and a DC gain of 40 dB is required, and a folded-cascode amplifier topology is used. Capacitor sizes can be very small (50 fF for both sampling and feedforward capacitor) because of the low resolution and matching requirements, reducing the power consumption. Since the signal swings are small and the resolution
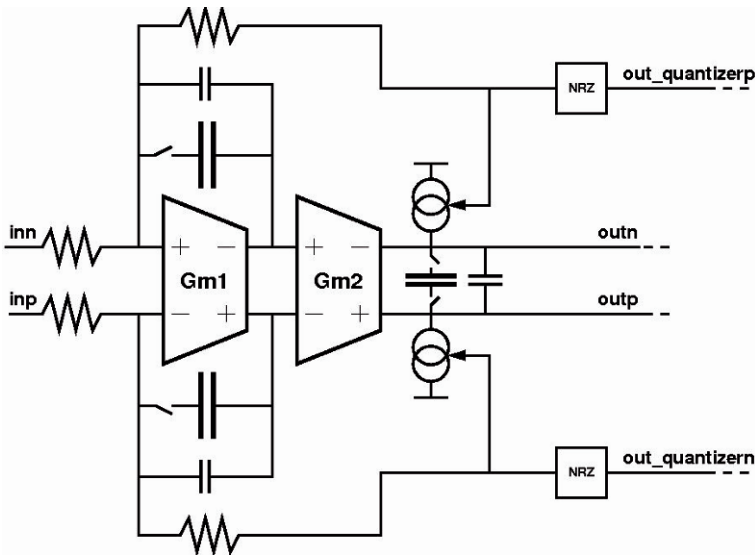
Figure 7.12.   Circuit schematic of the loop filter used in the reconfigurable A/D converter.

of the quantizer is moderate, simple NMOS switches suffice. No bootstrapped switches are required.

The $\Delta\Sigma$ modulator has a second-order continuous-time loop filter (see Fig. 7.12). The loop filter uses gm-C type integrators and switched current sources for feedback. Switchable capacitors are used to change the filter coefficients. Non-return-to-zero (NRZ) pulses are used to reduce jitter sensitivity. By ensuring fast enough switching, noise due to inter-symbol interference (ISI) is below the quantization noise floor. The DC gain requirement for the first OTA is 50 dB. Since signal swings are small throughout the convertor (0.15 Vptp), a one-stage amplifier can be used. A fully differential folded cascode topology is chosen for its relatively high DC gain and its ability to work at low power supplies (1.8V in this design). Gain-boosting techniques are not required, making the design more robust. The first OTA does not need source degeneration. If its GBW and DC gain are sufficiently large, the capacitive feedback will linearize the integrator. The OTA transconductance is 10 mS, while the input resistance is 800 Ω. For the second OTA however source degeneration is required. The transconductance of the input transistors is 1.5mS while the degeneration resistor is 8200 Ω. The source degeneration lowers the DC gain, but since the specifications of the second filter stage are relaxed, this doesn't pose any problems.

Reconfiguration or switching between the different operating modes of the converter is done by changing the capacitors in the loop filter, changing the sampling frequency and changing the oversampling ratio. The pipelined ADC

is only used in the wideband mode. In the other modes, only the quantizers in the first pipeline stage are used to create the 1-bit feedback signal, hence saving power. Table 2 shows the topology of the converter for the different modes. The loop-filter order remains the same in the different modes, only the OSR and quantizer resolution are changed. The multi-bit quantizer is only used in wideband mode (for WLAN).

A chip photograph is shown in Fig. 7.13. The active area is 1 mm2. The adaptive digital filter is not included on chip in this prototype implementation. The power consumption for the three modes is 37mW for the 20 MHz/58dB mode and 15mW for the 4MHz/60dB and 0.2MHz/70dB modes, respectively. The power consumption in the wide-bandwidth mode is quite comparable to dedicated solutions. The rather high power consumption in the lower-bandwidth

Table 7.2. Topology of the reconfigurable converter in the different modes.

mode	loop filter	OSR	quantizer
0.2 MHz/72dB (GSM)	2nd order CT	128	1b
4 MHz/60 dB (WCDMA)	2nd order CT	64	1b
20 MHz/60 dB (WLAN)	2nd order CT	12	6b (1b for loop)
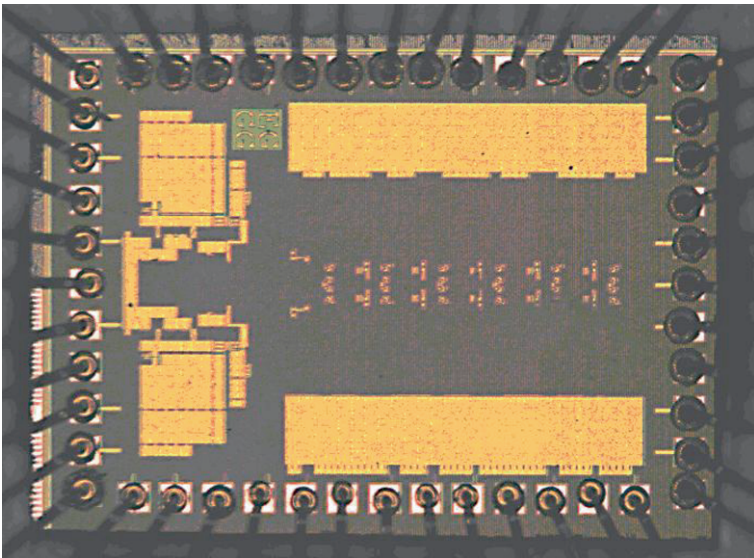
Figure 7.13. Chip photograph of a prototype multi-standard oversampling A/D converter in 0.18 μm CMOS technology.
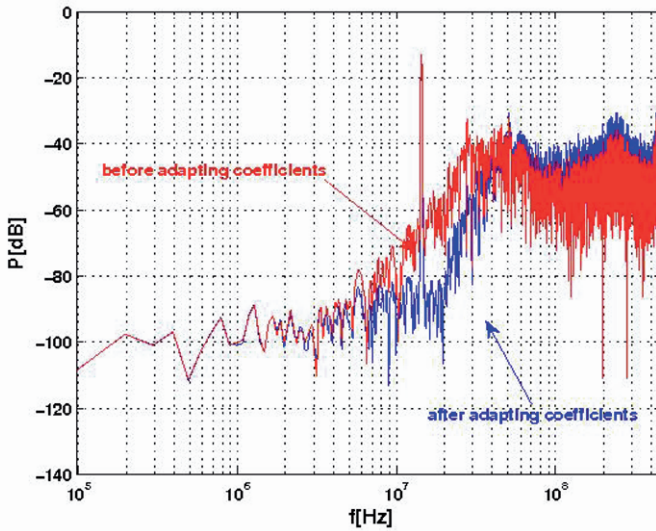
Figure 7.14. Simulated spectrum of the A/D converter before and after adapting the digital filter's tuning coefficients.

Table 7.3. Summary of results for the different modes of the prototype chip design.

Mode	Peak SNDR	Power Consumption
0.2 MHz	70 dB	15 mW
4 MHz	60 dB	15 mW
20 MHz	58 dB	37 mW

modes is due to the OTAs in the loop filter that were oversized for these modes (the reason is that in order to cut design time to meet the tape-out deadline the same OTAs were used in all the modes, whereas ideally scalable OTAs should be used that are scaled depending on the mode).

The simulated power spectrum of the wideband mode is shown in Fig. 7.14. The effect of adapting the digital filter's coefficients is clearly visible. The simulated peak SNDR in wideband mode is 58 dB with a power consumption of 37 mW. This includes both static and dynamic power consumption. The performance of the converter and the power consumption for the different modes is briefly summarized in Table 7.3. The distribution of the power consumption over the different building blocks in the wideband mode is shown in Table 7.4.

Table 7.4. Distribution of power consumption over the different blocks in the wideband mode.

Building block	Power consumption
loop filter stage 1	6 mW
loop filter stage 2	1.5 mW
track and hold	5 mW
OTA pipelined stage	4*3.25 mW
comparators	11*1 mW
total power	37 mW

These results indicate that it is feasible to implement multi-mode reconfigurable A/D converters in state-of-the-art CMOS technologies which can trade off bandwidth and dynamic range, at power levels that are competitive to dedicated solutions. The same is true for other reconfigurable analog and RF blocks as needed in software-defined reconfigurable front-ends for 4G wireless telecom systems.

4.4   Conclusions

This chapter has shown that reconfigurable software-defined radios are needed to cope with the flexibility and power requirements of 4G wireless systems. In order to produce a cost-effective 4G device that can handle multiple standards (cellular, WLAN, Bluetooth, DVB, etc.), flexible RF and analog baseband blocks are needed that can be reconfigured through digital programmable control. For the analog-to-digital converters in the receiver, this comes down to designing a power- and area-efficient reconfigurable converter with variable bandwidth and dynamic range. The requirements for such flexible multi-standard converter in 4G systems have been described. Oversampling converters are very suitable for this task. One of the main challenges however is to efficiently extend the bandwidth of these converters to accomodate WLAN standards (> 10 MHz bandwidth), which may require architectural combinations with pipelined or other Nyquist-rate converters.

This has been illustrated with the prototype design of a reconfigurable A/D converter that has three modes. The architecture chosen is a continuous-time $\Delta\Sigma$ A/D converter with a pipelined multi-bit quantizer and 1-bit feedback. Digital filter tuning is required to match the filter functions to suppress quantization noise. The chip has been realized in a 0.18 μm CMOS technology. The three different operating modes are: 20 MHz bandwidth/58dB SNDR, 4 MHz bandwidth/60dB SNDR, 0.2MHz bandwidth/70dB SNDR. The chip has an active area of 0.9 mm^2 and the power consumption for the most demanding mode (20

MHz/58 dB) is 37 mW. The ratio of the largest signal bandwidth to the lowest bandwidth is 100. This ratio is much larger than that of converters that are designed for cellular standards only (typically 20). Yet, the power consumption is quite competitive compared to dedicated solutions for the wideband mode.

Future work will focus on developing a reconfigurable A/D converter with even higher granularity of scalability, and with minimized power consumption in each mode.

References

[1] P. Kinget, M. Steyaert, Impact of transistor mismatch on the speed-accuracy-power trade-off of analog CMOS circuits, proceedings Custom Integrated Circuits Conference (CICC), pp. 333-336, 1996.

[2] A. Savla, Low-power design approaches for programmable-speed pipelined analog-to-digital converters, Master thesis Ohio State University, 2002.

[3] L. Breems, A cascaded continuous-time $\Delta\Sigma$ modulator with 67dB dynamic range in 10MHz bandwidth, proc. IEEE International Solid-State Circuits Conference, pp. 72-73, 2004.

[4] S. Patón, A. Di Giandomenico, L. Hernández, A. Wiesbauer, T. Pötscher, M. Clara, A 70-mW 300-MHz CMOS continuous-time $\Delta\Sigma$ ADC with 15-MHz bandwidth and 11 bits of resolution, IEEE Journal of Solid-State Circuits, Vol. 39, pp. 1056-1063, July 2004.

[5] A. Tabatabaei, K. Onodera, M. Zargari, H. Samavati, D. Su, A dual channel $\Delta\Sigma$ ADC with 40MHz aggregate signal bandwidth, proc. IEEE International Solid-State Circuits Conference, pp. 66-67, 2003.

[6] T. Burger, Q. Huang, A 13.5-mW 185-Msample/s $\Delta\Sigma$ modulator for UMTS/GSM dual-standard IF reception, IEEE Journal of Solid-State Circuits, Vol. 36, pp. 1868-1878, December 2001.

[7] A. Dezzani and E. Andre, A 1.2-V dual-mode WCDMA/GPRS $\Delta\Sigma$ modulator, proc. IEEE International Solid-State Circuits Conference, pp. 58-59, 2003.

[8] R. van Veldhoven, A tri-mode continuous-time $\Delta\Sigma$ modulator with switched-capacitor feedback DAC for a GSM-EDGE/CDMA2000/UMTS receiver, proc. IEEE International Solid-State Circuits Conference, pp. 60-61, 2003.

[9] K. Gulati and H.-S. Lee, A low-power reconfigurable analog-to-digital converter, IEEE Journal of Solid-State Circuits, Vol. 36, pp. 1900-1911, December 2001.

[10] Y. Geerts, High-performance CMOS Sigma-Delta converters, Ph.D. Dissertation Katholieke Universiteit Leuven, Belgium, May 2001.

[11] T. Leslie and B. Singh, An improved Sigma-Delta modulator architecture, proc. IEEE International Symposium on Circuits and Systems, pp. 372-375, 1990.

[12] T. Brooks, D. Robertson, D. Kelly, A. Del Muro, S. Harston, A cascaded Sigma-Delta pipeline A/D converter with 1.25 MHz signal bandwidth and 89 dB SNR, IEEE Journal of Solid-State Circuits, Vol. 32, pp. 1896-1906, December 1997.

[13] P. Kiss, J. Silva, A. Wiesbauer, T. Sun, U-K Moon, J. Stonick, G. Temes, Adaptive digital correction of analog errors in MASH ADCs - Part II. Correction using test-signal injection, IEEE Transactions on Circuits and Systems, part II, Vol. 47, pp. 629-638, July 2000.

PART III

RADIO DESIGN

Chapter 8

RECEIVER DESIGN FOR INTEGRATED MULTI-STANDARD WIRELESS RADIOS

Delia Rodríguez de Llera González, Ana Rusu and Mohammed Ismail

1.   Introduction

As we move beyond third generation (3G), the wireless scenario is rapidly becoming rather complex. A mobile device, be it a phone, a PDA or a notebook, is expected to be feature-rich, and able to work with several wireless standards while achieving the highest performance/price ratio. The existing wireless standards are very different from one another. Moreover, each of them is fragmented in different operating frequency bands both due to the limited spectrum availability and the particular regulations in different geographical areas. Hence, 4G systems need to provide multi-band multi-standard capabilities in order to be competitive. In the effort of providing the user with an always best connected experience, handhelds are to roam among these coexisting standards in a seamless manner. Thus, the system will adapt to the environment offering the best available quality of service for the different applications (data, voice, multimedia) the mobile terminal is running at a given time. This situation is highlighted in Figure 8.1, which shows examples of the various services available in different scenarios. Advances in both integrated circuits design and process technology permit the higher level of integration these systems require at a sufficiently low power consumption.

One of the main challenges for 4G wireless communications systems comes from trying to integrate the hardware of a number of existing wireless systems of newer generations that were conceived independently and were not meant to be integrated [1]. The requirements these various standards impose on the
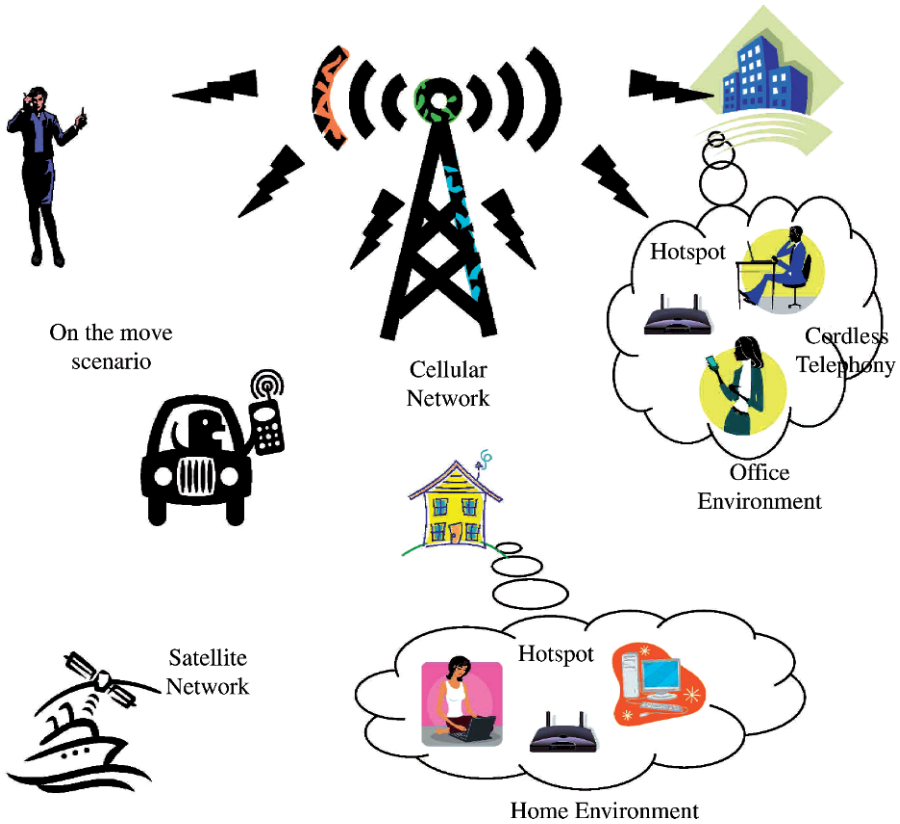
Figure 8.1.   Different wireless scenarios and the connectivity options they provide.

different blocks of the transceiver chain may vary immensely. Their optimal implementations may map into different architectures both at the system and the block levels. Building a receiver able to handle all these different standards is not an easy task. So why the effort? The answer comes in the form of product increased portability and reduced price, which are closely related to a reduction in silicon area and power consumption. One way of obtaining this is by finding a programmable architecture where most of the blocks can be reused for different standards [2].

   The receiver in a mobile terminal is one of the most challenging cases to solve. As opposed to the transmitter, the receiver's input signals lie in an uncontrolled environment full of interferers. In mobile terminals low power consumption is a must in order to ensure as long a battery life time as possible. High performance level is required in all operation modes.

Some of the major decisions to be made when designing a receiver are the architecture, the frequency translation scheme, the way to distribute the radio specs among the individual receiver blocks and the partitioning of the system. These decisions have to be made at a very early stage and will strongly determine the overall performance of the receiver as well as its cost and time-to-market. The higher the level a bad decision is made, the more time consuming and costly fixing the problem will be. Hence, a thorough study of the different possibilities is crucial at the system level in order to ensure that the market window for a product is not missed.

When aiming at a multi-standard receiver the problem is further complicated. A high level of performance hardware sharing is desired while keeping area and power consumption small. Flexibility has to be provided both at the architectural and block levels. It is no longer adequate or possible to tweak the design of the receiver blocks until they satisfy an optimal performance for a particular application. Being able to digitally tune and program the receiver at different levels is key in order to succeed in achieving the selectivity and sensitivity levels required by the ever changing target applications.

In contrast with their digital counterparts, analog and RF blocks have an extremely large design cycle. The level of uncertainty between simulated and fabricated circuits and the limitations of available automated design tools for analog circuits make the situation even worse. Intellectual Property (IP) block reuse is, therefore, very important not only due to the time saving it entails but also due to reliability issues. Including silicon proven blocks in a new design increases the chances of first pass success [3].

The realization of an efficient receiver budget and frequency planning is one of the most compelling problems RF engineers face nowadays. Even in the single standard case, the level of complexity of a wireless communications receiver is enormous. When the multi-standard case comes into the picture, this problem is aggravated with the need to share as much hardware as possible while keeping the performance levels high and the power consumption low.

The system level design is still nowadays done in many instances using the help of spreadsheets. Besides being error prone, this method is very limited in the number of different design possibilities it can explore within a given time. There is a number of EDA tools [4–9] that automate parts of this process. Most of them focus on analysis [6–8]. They may provide accurate models for the blocks [6, 8] or analyze the frequency behavior of certain parts of the circuit [9], but in general these tools provide little or no help at all to the RF engineer in the system level design process. Other reported tools [4, 5] and methodologies [10] help in the design process, but they only address the single-standard case.

The design of a multi-standard system is substantially more complex than the combination of the system level design of separate single-standard systems. Our multi-standard RF **T**ransceiver **A**rchitecture **C**omparison **T**ool, TACT, in-

troduced in [11] and thoroughly described in [12], is a tool that automates the process of design space exploration for multi-standard transceivers. This tool is aimed at easing the RF engineer's job as it fills a gap left by the already available CAD tools that address the transceiver design problem. For the time being, only the receiver side is implemented in the tool.

This chapter is organized as follows: Section 2 gives an overview of the general considerations to bear in mind when designing a receiver. Section 3 reviews some basic definitions and describes the method that allows mapping of the standard specs into receiver specs. Sections 4 and 5 introduce the frequency planning and budget design methodology proposed for, and implemented in, TACT. Section 6 shows a case study for a WCDMA/WLAN receiver design carried out using TACT.

2.   Multi-standard Receiver Design Considerations

The general considerations to follow when designing a multi-standard receiver could be summarized as follows:

1 Choose the target standards.

2 Obtain the standard specifications from the corresponding standardization body (IEEE, ETSI, etc.).

3 Develop a detailed list of requirements the receiver should meet based on the specs. Some of them have to be worked out based on the information given in the standard, others are given directly. These specifications include parameters such as LO's phase noise limits, required noise figure, second and third order intermodulation, filtering characteristics, etc. Section 3 gives detailed information on how to map the standard specs into receiver specs.

4 Compare the obtained receiver specs with common practice and update the specs if needed (you may not want to lag behind your competitors).

5 Choose the receiver architecture. For the multi-standard case, the receiver architecture should be as similar as possible for all the targeted standards in order to maximize the hardware sharing.

6 Develop a careful frequency plan that minimizes the effect of interferences along the receiver chain.

7 Design the receiver budget itself. This means distributing the gain, noise figure and intermodulation product levels among the receiver blocks. It is really the tricky part. It may require considerable time to work out all the numbers so that the specs for each of the blocks are practically realizable in a given process technology and the overall receiver meets the specs and is compliant with the test procedures specified in the standard. When targeting

a multi-standard receivers this gets more complicated. You always have to bear in mind that the specs for some of the blocks should be as similar as possible so that they could potentially be shared.

8 Follow the signal levels through the blocks up to the ADC for the maximum and minimum input signal, the blockers and the input thermal noise. This allows the calculation of the required dynamic range of the A/D converter, and therefore the necessary number of bits.

Usually several iterations between the last two steps are necessary in order to reach a fair distribution of the tough requirements. Both the multi-standard and the single standard cases are studied at the architectural and the block level. Basically, what has to be done once the specs are set is:

```
for all combinations of multi-standard and single-standard
    for all possible receiver architectures
        * find the frequency plan with less interferers
        * find a Rx budget that meets specs
         while not being tough on each block
    end;
end;
```

Overwhelming, isn't it? Each individual task is certainly easy to perform on its own, as we will see in the next sections. It is the number of times these operations have to be carried out, the endless possibilities of parameter distributions and the correlation between these parameters that makes this problem not only difficult but also complex.

Some degree of automation would certainly be appreciated by the RF systems engineer. The rest of the discussion will focus on general considerations related to receiver budget design as well as on-chip frequency planning. These issues will be addressed from the standpoint of our proposed solution implemented in TACT. TACT's simulation flow and interaction with the user are shown in Figure 8.2.

TACT consists of four interacting components: (1) a standard radio specifications to transceiver specifications mapping tool, (2) a pool of transceiver architecture models, (3) a pool of transceiver block models, and (4) a comparison tool, where the frequency planning and the receiver budget are performed. Figure 8.3 shows how they relate to each other.

3.   From Standard to Receiver Specs

This section will walk us through the procedure to follow in order to find the specs the receiver has to meet in order to be compliant with a certain standard.
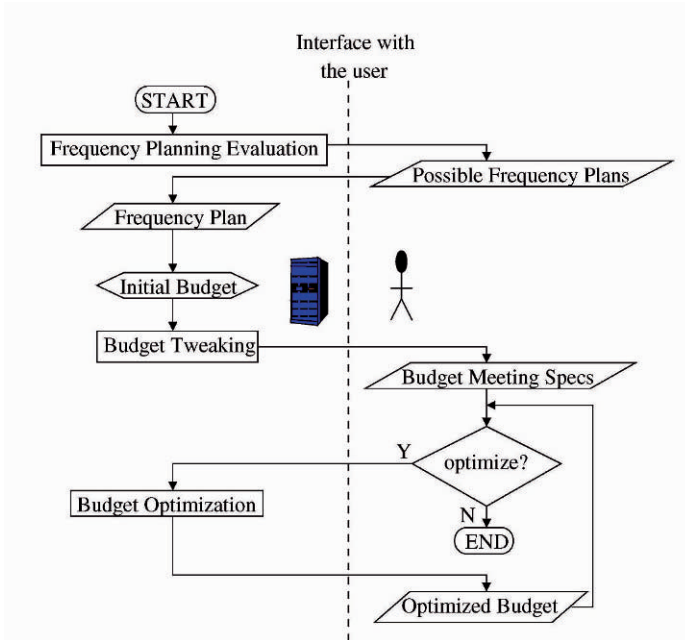
Figure 8.2. Simulation flow and interaction with the user of the frequency planning and budget tools.
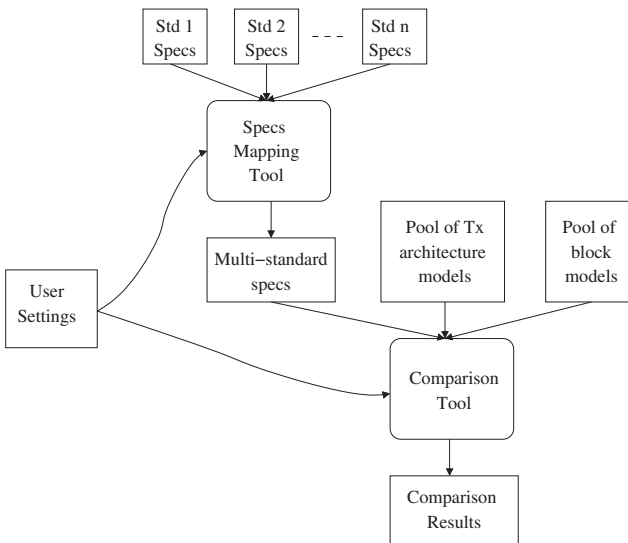
Figure 8.3. TACT components.

They comprise receiver noise figure, local oscillator phase noise, non-linearities and selectivity requirements. The parameters needed to compute them can be obtained in most cases from the radio specs of the standard. Some of them may not be specified for certain standards. Commonly used values for them are usually easy to find from commercial products.

3.1 Noise Figure

The noise figure of a system measures how much the signal to noise ratio (SNR) degrades when a signal passes through it [13]. The noise factor nf^1 of a system is given by

$$nf = \frac{SNR_{in}}{SNR_{out}} \qquad (8.1)$$

And the noise figure NF is the expression of the noise factor in dBs, that is:

$$NF = 10 \log nf \qquad (8.2)$$

The noise factor can also be expressed as:

$$nf = \frac{P_{in}/P_{RS}}{SNR_{out}} \qquad (8.3)$$

Where P_{in} is the input signal power and P_{RS} the source resistance noise power. Therefore,

$$P_{in} = P_{RS} \cdot nf \cdot SNR_{out} \qquad (8.4)$$

which integrated over the signal bandwidth gives the total mean square power of:

$$P_{in,tot} = P_{RS} \cdot nf \cdot SNR_{out} \cdot B \qquad (8.5)$$

for a flat channel where B is the channel bandwidth. Expressing this equation in decibel units we obtain:

$$P_{in,tot}|_{dBm} = P_{RS}|_{dBm/Hz} + NF|_{dB} + SNR_{out}|_{dB} + 10 \log B \qquad (8.6)$$

The source resistance noise power P_{RS} for a matched input at room temperature is given by:

$$P_{RS} = \frac{4kTRs}{4} \frac{1}{Rin}|_{R_s=R_{in}} = kT|_{T=290K} = -174dBm/Hz \qquad (8.7)$$

where $k = 1.3810^{-23}$ is the Boltzmann's constant.

Hence, Equation 8.6 can predict the system sensitivity $P_{in,min}$, that is, the minimum signal level that the system can detect with acceptable SNR, as:

$$P_{in,min}|_{dBm} = -174dBm/Hz + NF|_{dB} + 10\log B + SNR_{min}|_{dB} \quad (8.8)$$

where the sum of the first three terms is the total integrated noise or noise floor of the system. It immediately follows from this equation that:

$$NF|_{dB} = P_{in,min}|_{dBm} - SNR_{min}|_{dB} + 174dBm/Hz - 10\log B \quad (8.9)$$

Hence, the maximum noise figure a receiver chain is allowed to have can be calculated using the receiver sensitivity, the signal bandwidth and the minimum required signal to noise ratio at the output.

3.2   Local Oscillator Phase Noise

The phase noise of an oscillator at an offset frequency Δf of the carrier $\mathcal{L}(\Delta f)$ is given according to Leeson's model by:

$$\mathcal{L}(\Delta f)(dBc/Hz) = Ps(dBm) - Pb(\Delta f)(dBm) - 10\log B - SIR(dB)$$
$$(8.10)$$

where Ps denotes the input signal power, $Pb(\Delta f)$ the blocker power at an offset frequency Δf, and SIR the signal to interferer ratio. Depending on the standard, the blocker power at a certain offset frequency from the carrier $Pb(\Delta f)$ may be given by the blocking characteristics or by the adjacent channel characteristics. The one setting more stringent characteristics on the LO's phase noise performance should be taken.

The mechanisms originating phase noise are very complex and not fully understood [14]. Using Leeson's model for setting the local oscillator phase noise is an oversimplification, but it is a good starting point when setting the specs.

3.3   Non-linearities

Non-linearities of the analog/RF components may produce intermodulation products and/or harmonics of their input signals. These unwanted signals appear at the output of the receiver blocks along with the input signals. The intermodulation products fall at frequencies $f_{n,m}$ defined as:

$$f_{n,m} = \pm n \cdot f_1 \pm m \cdot f_2 \quad\quad\quad (8.11)$$

where n and m are integer numbers and f_1 and f_2 are the input frequencies. The order of the intermodulation product is $n + m$. In case of harmonic distortion Equation 8.11 is used with $m = 0$.
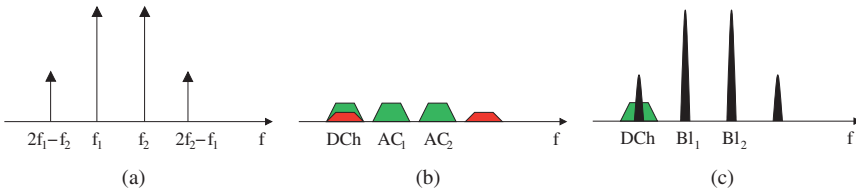
Figure 8.4. Third order intermodulation in a non-linear system 8.4(a). Two adjacent channels whose third order intermodulation product falls in the desired signal band 8.4(b). Two blockers whose third order intermodulation product falls within the desired signal band 8.4(c).

In a communications receiver, most of these IM products have very low power and/or are filtered out along the receiver chain. Nevertheless, some intermodulation products may fall close or within the desired signal frequency band and may have a power level large enough to significantly distort it.

Take for instance the third order intermodulation products depicted in Figure 8.4(a). Signals at frequencies f_1 and f_2 will produce third order intermodulation products at $2f_1 - f_2$ and $2f_2 - f_1$ when going through a non-linear system. Imagine now the situation shown in Figure 8.4(b). The desired channel (DCh) and two adjacent channels (AC_1 and AC_2 are present at the input of an RF component of the receiver chain. This is quite common since the channel selection is usually done at some intermediate frequency or at baseband. One of the third order intermodulation products originated by AC_1 and AC_2 will fall right on top of the desired channel producing a disturbance in the system. A similar effect will occur for any combination of blocker frequencies f_{b1} and f_{b2} such that

$$|f_{b1} - f_{b2}| = |f_{b1,b2} - f_{inband}| \tag{8.12}$$

Any such pair of blockers will produce a third order intermodulation product that will fall within the desired channel bandwidth at a frequency f_{inband}. This effect is illustrated in Figure 8.4(c).

Both the linearity of each block and the gain and filtering characteristics of the receiver chain will have an impact on the overall linearity of a receiver.

A means of comparing the linearity of different circuits is provided by the kth order intercept point (IPk). This point is located where the output level of the fundamental and the kth order IM product meet. The input power level corresponding to this point is the kth order input intercept point (IIPk). The output power level corresponding to this point is the kth order output intercept point (OIPk).

The power of the intermodulation products grows at a higher pace than the power of the signals that originate them as shown in Figure 8.5 when the devices work in weakly non-linear regions. This plot highlights also the fact that after a certain power level, the system starts saturating. This means that a linear
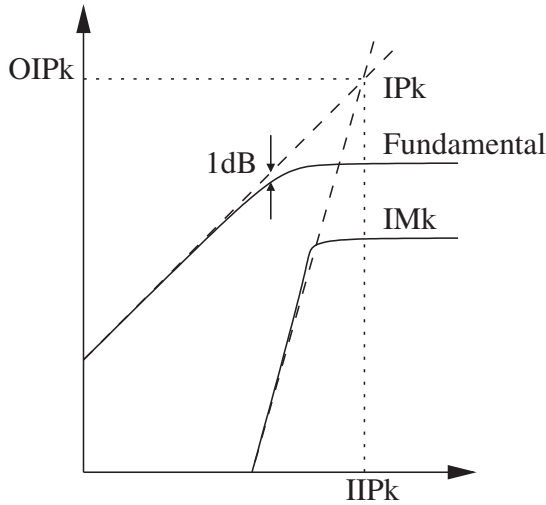
Figure 8.5. Calculation of the kth order intercept point (IPk).

increase in the input power level does not produce a linear increase increase in the output power level after a certain point. The 1-dB compression point, located where the actual curve is 1 dB away from the ideal one, is a measure of the saturation behaviour of a component or system. The IPk is, therefore, a mathematical construction more than a physical point that can be measured. It is calculated as the intersection point of the linear extrapolation of the fundamental and the kth intermodulation product lines. It is, nevertheless, a very useful construction. Since it is independent of the power of the input signals, it can be used to compare the linearity of different circuits as was mentioned before.

The power of an intermodulation product is inversely proportional to its order for the region where the system is weakly non-linear. Since the receiver will normally work in this region we will limit our discussion to second and third order non-linearities in this section. Higher order non-linearities could have a significant impact on the system in certain situations, though. They will be taken into account, at least qualitatively, when designing the frequency plan for the receiver.

It should be noted that the relative effect of odd and even intermodulation products depends on the chosen architecture. For instance, even order non-linearities are very detrimental in homodyne-like receivers whereas they are cancelled or nearly cancelled in most of the other receiver architectures as long as differential circuits are used.

Third Order Non-Linearity

The input third order intercept point (IIP3) can be calculated as:

$$IIP3|_{dBm} = P_{in}|_{dBm} + \Delta P/2|_{dB} \qquad (8.13)$$

Where P_{in} is the power of the input signal and ΔP is

$$\Delta P|_{dB} = P_{in}|_{dBm} - IM3|_{dBm} \qquad (8.14)$$

for a third order intermodulation product with a level of:

$$IM3|_{dBm} = P_{in,min}|_{dBm} - SNR_{min}|_{dB} - M|_{dB} \qquad (8.15)$$

where $P_{in,min}$ is the receiver sensitivity and a margin M is taken into consideration. These formulas can be used to calculate the minimum requirements a communications receiver should fulfill in terms of third order non-linearity.

Second Order Non-Linearity

The second order IM products $f_2 - f_1$ or $f_2 + f_1$ can become an issue in Low-IF and specially Zero-IF receivers since they fall in the low frequency band. In these types of receivers the second order IM products can be a more limiting factor than the third order ones.

The input second order intercept point (IIP2) can be calculated as:

$$IIP2|_{dBm} = P_{in}|_{dBm} + \Delta P|_{dB} \qquad (8.16)$$

Where P_{in} is the power of the input signal and ΔP is

$$\Delta P|_{dB} = P_{in}|_{dBm} - IM2|_{dBm} \qquad (8.17)$$

for a second order intermodulation product with a level of:

$$IM2|_{dBm} = P_{in,min}|_{dBm} - SNR_{min}|_{dB} - M|_{dB} \qquad (8.18)$$

where $P_{in,min}$ is the receiver sensitivity and a margin M is taken into consideration.

3.4   Selectivity Requirements

The selectivity requirements are determined by the analog-digital partitioning and the blocking and the adjacent channel selection characteristics given by the standard. Depending on the filter type (Butterworth, Elliptic, etc.) they may translate into different filter orders. The filter type choice should be done according to its pass-band ripple, stop-band attenuation and transition band so that not only the desired attenuation levels are reached but also the group delay, etc. The channel selection may be performed entirely in the analog domain thus relaxing the requirements of the ADC. When part or all of the filtering is moved to the digital domain the ADC specs become tougher. The amount of operations performed in the digital domain defines the analog-digital partitioning of the
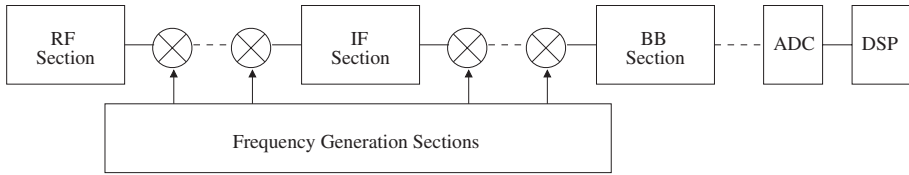
Figure 8.6. Abstract model of a generic receiver architecture. Not all the sections need to be present.

system. This partitioning has a strong impact on the system performance and requirements of the individual blocks.

4. Frequency Planning

Even though we are getting closer to the Software Defined Radio (SDR) paradigm, there are still a number of major practical problems associated with placing the ADC right after the antenna. Therefore, frequency translation in the analog domain is still a must in most receivers used in handheld devices at present. This, together with filtering and amplification stages eases the job of the analog-to-digital converter and keeps the power consumption sufficiently low to make the system practical for mobile terminals [15].

An interference oriented on-chip frequency planning is key in order to prevent the desensitization of the receiver as well as the interference of unwanted signals. These unwanted signals may appear in the different frequency bands where the desired channel falls during the different frequency transformations carried out in a receiver chain [13, 16, 17].

The objective of frequency planning is to find a frequency translation scheme with low distortion components, small limitations coming from the out-of-band blockers and lower filter order for maximum selectivity. The performance of the different possible intermediate frequencies (IF) can be measured through the levels of spurs falling in band as well as the bandwidths of the required filters. TACT takes into account not only the center frequencies of the signals, but also their bandwidths when performing all the computations [17].

4.1 Generic Receiver Architecture

Figure 8.6 shows an abstracted model that represents most of the architectures used in wireless communications receivers. In these receivers the RF input signal goes through a series of amplification, filtering and frequency translation stages until it is converted into a digital signal for digital post-processing.

Several factors have to be considered when choosing the set of frequencies and bandwidths for the successive frequency translation stages the signal goes
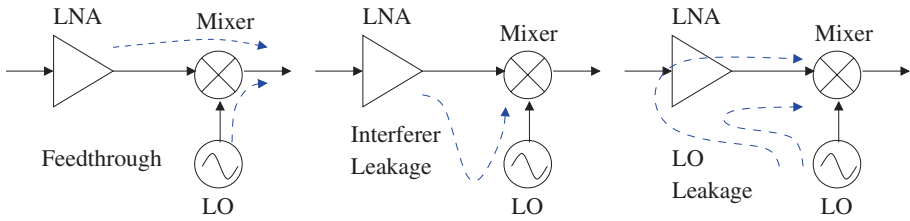
Figure 8.7. Signal feedthrough and self-mixing.

through [16, 13]. The compromise made between the filtering characteristics, the linearity of the components of the receiver chain, and the ADC dynamic range will depend on the particular application and the availability of already designed IPs.

Unwanted signals may have different origins, namely: channels using the same frequency at the same time in neighbouring cells (FDMA systems) or in the same cell with a different code (CDMA systems), channels located in adjacent frequencies, out of band blockers present in the wireless environment, other interfering signals present on-chip coming from other parts of the circuit (signals coming from the transmitting side that appear in the receiving side through on-chip coupling mechanisms, for instance), etc.

Non-linearities of the analog/RF components may produce intermodulation products and/or harmonics of their input frequencies as was explained in Section 3.3.

4.2 Mixing and Frequency Translation

The mixing process itself takes advantage of this non-linear behaviour, so undesired in other blocks of the receiver chain, in order to perform the frequency translation along the receiver chain. Mixers are not ideal components either. Due to their finite input-output isolation, an attenuated version of the input signals appear at the output. This phenomenon is known as signal feed-trough and is illustrated in Figure 8.7. It could have detrimental consequences in the system if these fed-through signals experience further non-linear processes that place them within the signal band. Another problem, acute in the case of zero-IF receivers, is self-mixing of the local oscillator and/or interfering signals. These signals, which might have a high power level, will leak under certain conditions to the other input terminal of the mixer as shown in Figure 8.7, and mix with themselves falling within the desired channel band. Different receiver topologies suffer from different problems related to their particular characteristics. These effects, described thoroughly in [13, 16, 18], are modeled in TACT [11, 17].

The set of parameters that have to be determined when carrying out the on-chip frequency planning are:

- The number of intermediate frequencies needed.

- The center frequency of the intermediate frequency band(s).

- The bandwidth of the filter(s) at RF, IF and BB.

The first step is to determine the number of recommended translations. The aim of performing more than one frequency translation is to obtain a good selectivity while keeping the requirements of the filters as relaxed as possible. The method described in [16] is used in TACT.

The next step is to determine the range of possible intermediate frequencies. There is a number of conditions the intermediate frequency band has to meet [16], namely: the intermediate frequency band should not overlap with the signal band or the local oscillator, the relative bandwidth of the filters should make them feasible, and the image band should be rejected when the receiver architecture is sensitive to it.

These conditions are translated into different equations depending on the relative location of the signal RF band, the local oscillator and the IF band. This leads to different intermediate frequency ranges. An extended version of the equations proposed in [16] is used in TACT.

The tool also displays commonly used frequency values within the intermediate frequency range [16]. This can be a valuable piece of information when using commercial-off-the-shelf (COTS) components. The use of standard components, which are produced in large volumes by several vendors, can have a positive impact on the cost.

4.3   Intermediate Frequency Search

Once the intermediate frequency ranges are determined for all the signal bands the evaluation of the different possibilities starts. These signal bands can belong either to different or the same standard and may or may not overlap.

The user can set the number of intermediate frequency values N to be evaluated. Thus, the granularity of the frequency search space is determined. These points are distributed along the different IF ranges in such a way that in the frequency intervals common to two or more of them the evaluation is performed exactly for the same intermediate frequency values. This eases the evaluation of the multi-standard case.

The evaluation of the performance of each IF with respect to the interferers is a two step process. First the interferences coming from signals related to the given standard are considered. Then, the effect of out-of-band blockers is analyzed.

The bandwidth ranges for the RF and IF filters are initially set so that:

1 Their relative bandwidths (B_{RF}/f_{RF}, B_{IF}/f_{IF}) make them feasible.

2 They are larger than the channel bandwidth $B_{RF}, B_{IF} > B_{ch}$.

3 They are smaller than the channel separation for the last IF filter $B_{IF} < B_{sep}$. In the case of the RF filter bandwidth, the filtering of the image band has to be ensured. Therefore $B_{RF} < 2f_{IFmin}$.

As mentioned before, the signals related to the standard under analysis are considered first. These are signals whose characteristics are known a priori. They comprise the different input channels, their corresponding local oscillator frequencies and a number of other signals present on-chip that may cause disturbances due to on-chip coupling. These signals may be the output of the power amplifier of a transmitter sitting on the same die, the voltage controlled oscillator (VCO) frequency, any digital clock present on chip, etc.

4.4 Intermodulation and Harmonics

All the frequency bands of the undesired intermodulation products, harmonics, self-mixing products and mixer feed-through signals coming from the combination of these "known" signals are computed. Should any of these signals fall within the signal band, either at the RF or the IF frequency bands, the overlapping range, m and n in Equation 8.11, and the frequencies of the signals that originate that distortion component are stored in TACT for further display to the user both in numerical and graphical form. At this stage the actual power levels of the signals at every point of the receiver chain are not known, but the origin of the distortion component and the order give clues to an RF designer about the power level and final impact of these signals in the receiver performance.

The minimum (most detrimental) order of the distortion components falling within each of the intermediate frequency bands for the standards under consideration is displayed to the user. Section 6 shows an example of such plot. When looking at a single intermediate frequency, all the distortion components with their order, overlapping range with the signal band and origin are identified.

4.5 Out-of-band Blockers

When evaluating the effect of out-of-band blockers that fall within the RF filter pass-band a different method from the one discussed in the previous section is implemented. The characteristics of the out-of-band signals are not known. The national or international regulatory bodies of the radio spectrum may enforce a maximum power level for signals close in frequency to a licensed band. However, no information on the bandwidth or dynamics of potential interferers are known.
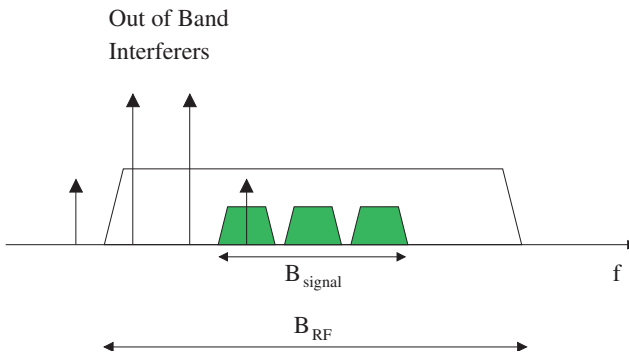
Figure 8.8.   Effect of out-of-band interferers after going through a non-linear stage.

It is well known that any combination of blocker frequencies f_{b1} and f_{b2} satisfying Equation 8.12 will produce third order intermodulation products that will fall in band at a frequency f_{inband}. This effect, illustrated in Figure 8.8 cannot be avoided. Its impact will depend on the linearity of the receiver blocks. Increasing the linearity of these components and/or reducing the RF filter bandwidth mitigates the impact of these effects, but changing these parameters is not always possible.

When an out-of-band blocker and one of the "known" signals mentioned above are involved in the non-linear process, equation 8.11 cannot be used directly. Nevertheless, the same principle can be applied from a different standpoint. This process, illustrated in Figure 8.9, unveils the out-of-band frequency regions that may produce disturbances in the system. This information provides a deeper insight into the performance of each of the evaluated intermediate frequencies. It may change either the frequency plan choice or the requirements of the filters for some of the intermediate frequencies. These undesired blocker bands are computed as follows:

- The combinations $|f_i \pm n \cdot f_{o/ch}|$ are calculated.

- The combinations $m \cdot f_{blockerband}$ are calculated. The blocker bands are the RF filter pass-band without the signal bands.

- The overlap between the resulting bands is computed.

- The blocker bands corresponding to the overlapping range are computed.

5.   Receiver Budget

Many tradeoffs have to be made when fixing the characteristics of each of the blocks. The interdependency between the overall noise and non-linearity

Figure 8.9. Calculation of the most detrimental out-of-band interfering frequency bands.

characteristics on the gain and selectivity of all the blocks makes this task difficult.

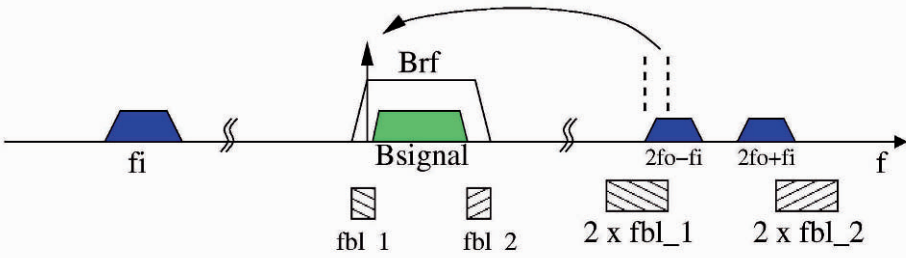The receiver budget is realized taking the radio specifications of the standards under consideration, the receiver architectures a user wants to explore and the high level models of the receiver blocks. At the output comes a set of candidate receiver budgets with their performances shown to the user. The cost functions that are evaluated at this stage are the overall noise figure (NF), the second and third order intercept points (IIP2, IIP3), and the number of bits of the analog-to-digital converter (ADC).

This parameter distribution should be done in such a way that:

- The specs for each block are practically reasonable.

- The aggregate parameters computed meet the specs determined using the procedure described in Section 3.

We will see how to check this last condition in the coming sections. As for setting reasonable specs for each of the blocks, there is no common procedure or methodology that can be used. It is based on rules of thumb, experience, contact with experts and a lot of digging in published work. This is a key issue without an easy answer.

The architecture selection and the order in which the amplifying and filtering operations are performed [13] will lead to different block combinations and therefore, different results.

Although current trends call for implementation of different standards into one common radio architecture [19] using programmable blocks, this may not be feasible and may come at the expense of a degraded overall performance of the receiver chain. As a consequence, one may find it better to duplicate a block, rather than using a programmable common block. Even two completely separate signal paths might be a better option. Hence, both the multi-standard and the single standard cases are studied at the architectural and the block level as was mentioned in Section 2.

5.1   Sensitivity and Gain Levels

The gain budget is done both for the sensitivity value (maximum gain option) and the maximum input power signal (minimum gain option). The automatic gain control (AGC) tuning range is determined by the difference between the minimum and the maximum gain levels. Thus, the far-near problem can be addressed leading to an overall power save and avoidance of undesired effects such as desensitization.

The gain distribution should be done bearing in mind:

- The signal levels of the standards (maximum, minimum, noise floor, blockers, etc.).

- Reasonable gain/loss levels for all the blocks (with some margin for circuit non-idealities).

- The gain range for the programmable gain stages (LNAs, VGAs) so that the dynamic range requirements for the ADC are as relaxed as possible for the chosen analog-digital partitioning.

- The impact of the gain in the agreggate NF, IIP3 and IIP2.

The aggregate gain for the maximum and minimum gain options is calculated by the addition of the individual gains in dB.

5.2   Noise Figure and Linearity

The overall noise factor for the cascaded receiver blocks can be calculated using Friis equation:

$$nf = 1 + (nf_1 - 1) + \frac{nf_2 - 1}{A_1} + \cdots + \frac{nf_m - 1}{A_1 \cdots A_{(m-1)}} \qquad (8.19)$$

where nf_i is the noise factor and A_i is the power gain of the i-th block. The noise figure is the equivalent in dB of the noise factor, $NF = 10 log nf$.

This equation is only valid for cascaded components where no frequency translation is performed [16]. When frequency translation is carried out, as is the case in most receivers, the contribution of the image noise band should be considered too. The characteristics of most receiver blocks in this image noise band is very difficult to predict at this early stage of tough.

As Equation 8.19 clearly shows, the overall noise figure will be determined by the first stages of the cascaded receiver. Hence, front end blocks with small NF and high gain are crucial for a receiver meeting the NF specs.

It should be noted as well that lossy blocks will amplify the noise contribution of the stages following them.

As far as linearity is concerned, the third order intercept point at a given frequency is equal to:

$$\frac{1}{IP_3^2} = \frac{1}{IP_{3,1}^2} + \frac{A_1}{IP_{3,2}^2} + \cdots + \frac{A_1 \cdots A_{n-1}}{IP_{3,n}^2} \qquad (8.20)$$

where $IP_{3,i}$ is the third order intercept point and A_i is the gain of the i-th block. The second order intercept point has an equivalent equation.

It can be seen by inspection of Equations 8.19 and 8.20 that qualitatively, linearity and noise performance impose opposed conditions to the gain distribution. In terms of noise figure a high front-end gain is desirable whereas in terms of linearity a low front-end gain is desirable. A fairly small front-end gain makes intuitive sense when thinking of all the blockers entering the receiver front-end together with the desired signal. Having a high front-end gain before any filtering of these undesired signals might be very detrimental for the the performance of the system if the RF front-end components are not extremely linear and have a high dynamic range.

5.3   ADC's Dynamic Range and Effective Number of Bits

The number of bits of the ADC is related with the dynamic range (DR) of its input:

$$DR_{ADC} = P_{max} - P_{noise} + M \qquad (8.21)$$

where P_{max} is the maximum signal power present at the ADC input, P_{noise} is the input noise floor, and M a margin set by the user. The effective number of bits of the ADC $ENOB_{ADC}$ can be calculated as:

$$ENOB_{ADC} = \frac{DR_{ADC} - 1.76}{6.02} \qquad (8.22)$$

5.4   Receiver Budget Optimization in TACT

Finding a parameter distribution that meets specs is the first goal the algorithm implemented in TACT focuses on. Then, both the solution and the room for improvement for each of the parameters is presented to the user. Further optimization is performed using basically the same algorithm as when finding the initial distribution that meets specs. This algorithm works as follows:

The loop of finding a solution that meets the specs is entered. First, whether or not the specs are achievable at all is checked. If any of the cost functions is not achievable, execution is halted, the best possible solution is displayed, and the user is asked to revise the parameter margins. Otherwise execution continues.

```
while specs_met == false
   if gain_redistrib_needed == true
        redistribute_gain;
   else % gain redistribution not needed
     if rand < p % change it anyway sometimes
          redistribute_gain;
     else
          redistribute_params_not_meeting_specs;
     end;
   end;
   cost = check_specs;
   if cost < specs
        specs_met = true;
   end;
   if cost < best_cost
        best_budget = this_budget;
   end;
end;
```

Figure 8.10.   Algorithm to find a budget meeting specs.

The first step is to generate a seed solution. Each parameter of each block is set to the value that makes its specs most relaxed. Then, the operations sketched in Figure 8.10 are performed.

The routine that changes the gain distribution has to determine both the direction of the change (increase or decrease the gain), the location of the change (front-end, back-end, everywhere) and the amount of gain change to be introduced. When changing the gain, it is ensured that the new gain value is within the range limits. The absolute value of the gain variation is random within these limits.

In a receiver chain, there are blocks with variable gain and blocks with fixed gain. When the multi-standard case is being evaluated, the gain changes for the blocks programmable in gain are introduced with probability one. If the block is not programmable in gain, the gain reassignment is taken into consideration with probability p, where p is decided by the user. Depending on the gain values the various standards try to assign to a block and the direction they are trying to push the new value to, a different gain value will be resolved.

Table 8.1. Execution time results.

Benchmark	$t_{min}(s)$	$t_{max}(s)$	$t_{mean}(s)$	σ
WCDMA/WLAN	3.07	25.47	9.69	3.73

The reassignment of the values of the noise figure and intercept points for each of the blocks is done according to their impact on the overall system. It also takes into account how much margin for changing the parameter has. This parameter reassignment only takes place when the overall value does not meet specs.

If the ENOB specs are not met, the minimum gain of the AGC is adjusted, the filtering specs are hardened or both. The probability of choosing each of these options depends on the origin of the signals that determine the ADC dynamic range.

The execution time of these algorithms depends on many factors. When doing the frequency plan, the order of non-linearities considered and the number of intermediate frequency points are the most important factors. The budget optimization is based on simulated annealing [20]. Its execution time is non-deterministic and very dependent on the cooling temperature dr:annealing and the goal functions. Table 8.1 shows the maximum, minimum, mean and standard deviation of the execution time for the specs shown in Table 8.5. The statistics are calculated over 200 runs of the tool.

Once the receiver specs have been set, tools providing more accurate models, such as ADS, can be used so that second order effects can be accounted for. Final adjustments in the block specs may have to be done at this point. It is recommended to provide some margin in the cost functions when doing the budget design using TACT in order to minimize the number of design iterations. An interface with ADS is under construction. This interface will provide a bridge between high and low level receiver design.

6. Case Study:WCDMA/WLAN Receiver Budget

This section presents a case study of a WCDMA/WLAN multi-standard receiver. The radio characteristics of these two standards are summarized in Table 8.2.

Zero-IF is one of the most commonly used receiver architectures for multi-standard applications. TACT provides though the possibility of exploring intermediate frequencies different from zero. For $f_i \neq 0$, choosing a high-side LO makes the recommended intermediate frequency range of WCDMA go from 24.02 MHz to 500 MHz and the one of WLAN from 40 MHz to 2 GHz. This

Table 8.2.   Summary of the WCDMA (TDD) and WLAN(802.11b) RF specifications.

Parameter	WCDMA		WLAN	
RF Frequency Band	2010-2025	MHz	2400-2485	MHz
	1900-1910	MHz		
RF Channel Bandwidth	3.84	MHz	20	MHz
Channel Separation	5	MHz	5/25	MHz
Sensitivity	-117	dBm	-76	dBm
Max Power Level	-25	dBm	-10	dBm
Adjacent Ch. Selectivity	33	dB	35	dB

range can be modified by the user. Some of the commonly used frequencies within this range are 140, 190 and 380 MHz, which are widely used in UHF and microwave broadband receivers [16]. A user building a commercial off-the-shelf (COTS) components based system might find this information useful.

Moreover, in the WCDMA mode, for an intermediate frequency of 44.3 MHz the IF filter bandwidth B_{IF} ranges between 3.84 and 5 MHz. For a higher intermediate frequency such as 417.6 MHz where the relative bandwidth is narrower, 4.2 MHz $< B_{IF} < 5$ MHz.

Figure 8.11 shows one of the outputs provided by the frequency planning tool. This plot gives a snapshot of the intermodulation and harmonic content due to in-band interferers for different intermediate frequencies. The y-axes shows the lower order of the distortion components falling within each of the intermediate frequency bands. For a given pair of input signals, distortion components have larger power the smaller their order is. Hence, low order distortion components are, in general, more detrimental to the system. However, the order is not the only important factor as was mentioned before. The impact of odd and even distortion components depends on the receiver architecture [13, 16]. Their cancellation or effect is usually addressed in different ways. The origin of the signals producing distortion components and the frequency overlap with the desired signal's bandwidth are also key in determining the performance of each intermediate frequency. TACT stores this information and makes it available to the user when only one intermediate frequency is selected.

Let us continue with the example using WCDMA with an IF of 100 MHz. Taking the worst case RF bandwidth from the interference point of view (the largest bandwidth) the left-hand side blocker band goes from 1919.9 MHz to 2010 MHz. The band corresponding to twice a blocker band is located between 33839.9 and 4020 MHz. One of the intermodulation products coming from the combinations $|f_i \pm n \cdot f_{o/ch}|$ that is located within that range, is the one
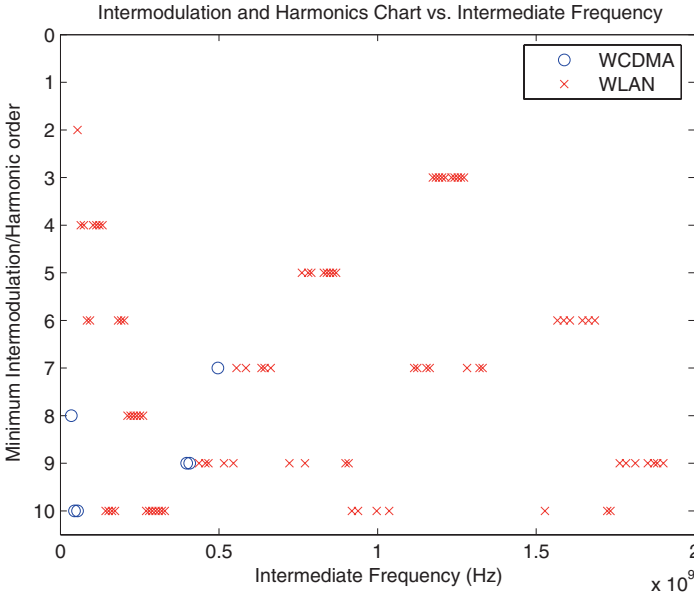
Figure 8.11. Minimum order of distortion components vs. the intermediate frequency band within which they fall.

that covers the 3922.4-3927.4 MHz band and corresponds to n = 2 for the first channel of the signal band and an f_i of 100.12 MHz. This corresponds to the blocker band 1961.2-1963.7 MHz. Therefore, should an interferer be located within this band, a fourth order intermodulation product between the first signal channel and that blocker would fall within the intermediate frequency range centered at 100.12 MHz.

Once the user has chosen a frequency plan, it is time to move on to designing the receiver budget. We have chosen a zero-IF receiver architecture for this case study since it is one of the preferred architectures for multi-standard systems. The RF standard specs (summarized in Table 8.2) are mapped into receiver specs by TACT. The resulting receiver specs are shown in Table 8.3.

In order to find a parameter distribution meeting these receiver specs using the zero-IF receiver shown in Figure 8.12, the budget tool within TACT is executed. The distribution of the gain, noise figure, linearity performance, and filtering characteristics changes along with the simulation step. These parameters are readjusted in order to meet the noise figure and linearity specs set by the standard, the ADC dynamic range specified by the user and the analog-digital partitioning that results from them. The tool is asked to meet the requirements shown in Table 8.3 and with a maximum dynamic range of 60 dB for WCDMA and of 40 dB for WLAN.

Table 8.3.   Summary of the receiver specs for WCDMA (TDD) and WLAN (802.11b).

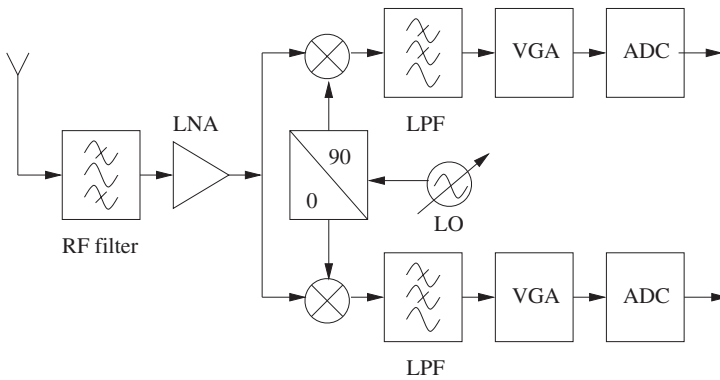| Parameter	WCDMA		WLAN	
NF	9	dB	11	dB
IIP3	-17	dBm	-5	dBm
IIP2	14	dBm	23	dBm
Phase Noise	-120	dBc/Hz@5MHz	-123	dBc/Hz@5MHz
	-126	dBc/Hz@10MHz		
	-139	dBc/Hz@15MHz		
	-137	dBc/Hz@20MHz		

Figure 8.12.   Zero IF receiver architecture.

Figure 8.13 shows the resulting signal levels along the receiver blocks. The levels of the desired input signals (maximum and minimum option), adjacent channels, blockers and noise floor are plotted against the different blocks. These signal levels are shown at the input of the RF filter (1), the LNA (2), the mixer (3), the baseband filter (4), the baseband VGA (5), and the ADC (6). The redistribution of the block characteristics performed during the optimization of the budget makes these signals evolve with the simulation step until they reach the levels shown in this figure. The optimization is done considering the multi-standard case in order to ease the hardware sharing of the receiver blocks.

The final parameter distribution obtained in this run of the budget tool is shown in Table 8.4. The performance of the proposed receiver is summarized in Table 8.5. The obtained performance is shown to meet or exceed the requirements of the WCDMA/WLAN standards. As such, the case study validates the benefits of the proposed tool.
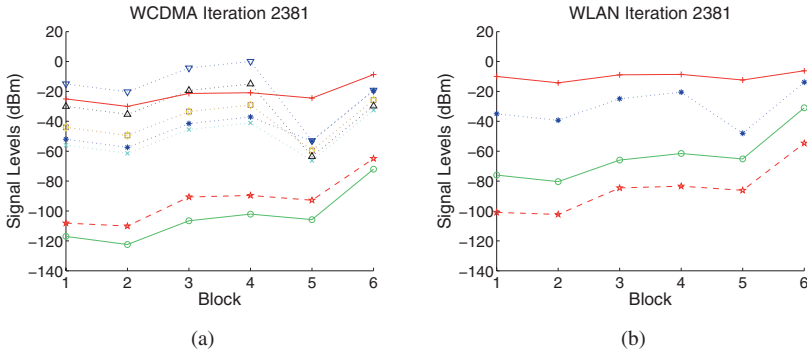
Figure 8.13. Signal levels along the blocks for WCDMA and WLAN for a typical TACT run. The WCDMA signals shown in 8.13(a) correspond to: + P_{max}, o P_{min}, * Adjacent Channel at 5 MHz offset, x Adjacent Channel at 10 MHz offset, □ Adjacent Channel at 15 MHz offset, ◇ Out-of-band Blocker at 15 MHz offset, ▽ Out-of-band Blocker at 60 MHz offset, △ Out-of-band Blocker at 85 MHz offset, ★ Noise Floor. The WLAN signals shown in 8.13(b) correspond to: + P_{max}, o P_{min}, * Adjacent Channel at 25 MHz offset, ★ Noise Floor

Table 8.4. Parameter distribution for the proposed WCDMA/WLAN multi-standard receiver.

WCDMA					
Parameter	RF filter	LNA	Mixer	BB filter	VGA
Gain (dB)	-3.7429	15.5271	4.2427	-3.5662	33.0083
NF (dB)	2.5758	2.2830	5.8957	9.7592	10.2608
IIP3 (dBm)	3.2397	-0.3067	1.6909	3.2397	4.3840
IIP2 (dBm)	40.0000	35.0000	55.0000	45.0000	50.0000
WLAN					
Parameter	RF filter	LNA	Mixer	BB filter	VGA
Gain (dB)	-3.7429	14.9321	4.2427	-3.5662	35.5502
NF (dB)	3.1739	2.5971	6.7652	12.6336	12.8973
IIP3 (dBm)	13.4248	11.0595	12.4423	13.4248	13.7341
IIP2 (dBm)	40.0000	35.0000	55.0000	45.0000	50.0000

7. Conclusions

This chapter addressed the receiver budget problem with special focus on emerging multi-band,multi-standard fully integrated radios. The discussions were given in the context of a Transceiver Architecture Comparison Tool, TACT, that automates the process of design space exploration for multi-standard radio transceivers. First an overview of design consideration and requirements

Table 8.5. Specifications and performance of a typical run for a WCDMA/WLAN multi-standard receiver.

Standard	WCDMA		WLAN	
Parameter	Specs	Result	Specs	Result
$DR_{ADC}(dBm)$	60	58.3	49	48.9
$ENOB_{ADC}$	10	9.3	8	7.8
$Gain(dB)$	-	45.5	-	47.4
$NF(dB)$	9	6.4	11	7.7
$IIP3(dBm)$	-17	-15.6	-5	-4.8
$IIP2(dBm)$	14	27.6	23	28.1

is given, followed by a description of a method that does mapping of standard(s) specs onto receiver specs. Finally,the frequency planning algorithms as well as budget design and optimization schemes used within TACT are described and demonstrated in a case study of a programmable multi-standard WCDMA/802.11b direct conversion receiver for fourth generation (4G) wireless.

Notes

1 In this chapter the noise factor of a system is abbreviated as nf instead of f (the most extended notation) in order to avoid confusion with frequency, which is referred to as f.

References

[1] Xiaopeng Li and Mohammed Ismail. *Multi-standard CMOS Wireless Receivers. Analysis and Design*. Kluwer Academic Publishers, 2002.

[2] Pieter Hooijmans. Architectures for mobile RF convergence and future RF transparency. *RFDESIGN*, February 2006.

[3] James Wilson and Mohammed Ismail. *Radio Design in Nanometer Technologies*, chapter Design Techniques for First Pass Silicon Success. Springer, 2006.

[4] Jan Crols, Stéphan Donnay, Michiel Steyaert, and Georges Gielen. A High-Level Design and Optimization Tool. In *Digest of Technical Papers. IEEE/ACM International Conference on Computer-Aided Design*, 1995.

[5] Georges G. E. Gielen. Modeling and Analysis Techniques for System-Level Architectural Design of Telecom Front-Ends. *IEEE Transactions on Microwave Theory and Techniques*, 50(1), January 2002.

[6] http://www.eesof.tm.agilent.com/. *Advance Design System EDA Software (ADS)*.

[7] http://www.nathaniyer.com/. *Cascade*.

[8] http://www.ansoft.com/. *ANSOFT*.

[9] http://www.eagleware.com/products/genesys/ whatif.html/. *Eagleware Elanix WhatIF*.

[10] Sheng Wenjun and Edgar Sánchez-Sinencio. System level design of radio frequency receivers for wireless communications. In *ASIC, 2003*.

[11] Delia Rodríguez de Llera González, Ana Rusu, Mohammed Ismail, and Hannu Tenhunen. TACT: A Multi-standard RF Transceiver Architecture Comparison Tool. In *IEEE Midwest Symposium on Circuits and Systems (MWSCAS)*, 2005.

[12] Delia Rodríguez de Llera González. *Automatic Design Space Exploration of Integrated Multi-Standard Wireless Radio Receivers* . Licenciate thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, 2006.

[13] Behzad Razavi. *RF Microelectronics*. Prentice Hall PTR, 1998.

[14] Waleed Khalil and Bertan Bakkaloglu. *Radio Design in Nanometer Technologies*, chapter Challenges in the Design of PLLs in Deep-Submicron Technology.

[15] Christer Svensson and Stefan Andersson. *Radio Design in Nanometer Technologies*, chapter Software Defined Radio - Visions, Challenges and Solutions. Springer, 2006.

[16] Manuel Sierra Pérez, Belén Galocha Irangüen, José Luis Fernández Jambrina, and Manuel Sierra Castañer. *Electrónica de Comunicaciones*. Pearson, Prentice Hall, 2003.

[17] Delia Rodríguez de Llera González, Ana Rusu, and Mohammed Ismail. A Frequency Plan Evaluation Tool for Multi-standard Wireless Transceivers. In *IEEE Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*.

[18] Sven Mattisson. *Radio Design in Nanometer Technologies*, chapter Cellular RF Requirements and Integration Trends. Springer, 2006.

[19] Rami Ahola and et al. A Single-Chip CMOS Transceiver for 802.11a/b/g Wireless LANs. *IEEE Journal of Solid-State Circuits*, 39(12), December 2004.

[20] P.J. van Laarhoven and E.H. Aarts. *Simulated Annealing: Theory and Applications*. Kluwer Academic Publishers, 1987.

Chapter 9

ON-CHIP ESD PROTECTION FOR RFICS

Elyse Rosenbaum and Sami Hyvonen

1.   Introduction

Electrostatic discharge, or ESD, has been referred to as a "pervasive reliability concern" for ICs [1] because there are so many different scenarios under which an integrated circuit may lie in the path of a static discharge. For example, a person may acquire static charge by walking across a carpet and then dissipate this charge by picking up an IC and plugging it into a grounded socket. The IC may itself acquire static charge, for example, by sliding down a plastic shipping tube, and this charge will dissipate when the IC first contacts a conductor. These events would be described by the Human Body Model and the Charged Device Model, respectively [2]. Human Body Model type events are characterized by a risetime of about 10 ns, duration of about 150 ns, and peak current on the order of a few amperes. Charged Device Model type events are characterized by a risetime of about 250 ps, duration of 1-2 ns, and a peak current of about 10 A. RF transistors are designed to operate at milliamp current levels, and if the ESD current were to pass through one of these devices, it would be destroyed, leading to circuit failure. Therefore, between any arbitrary pair of pins, a safe path must be provided for dissipation of static charge. This is achieved by on-chip ESD protection circuits.

2.   Full-Chip Protection Topology

A topology for full-chip ESD protection is illustrated in Fig. 9.1 for a chip with multiple power supply domains. Other topologies are possible, but this one is probably the most robust [3]. The ESD bus should be a continuous
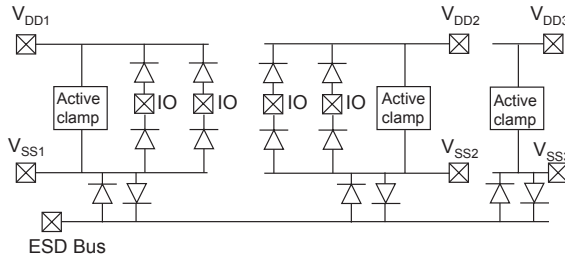
Figure 9.1. Representative drawing of a full-chip ESD protection network. Each I/O pad is protected by a dual-diode circuit.

ring around the chip [3]. If it is a dedicated ESD bus, then it need not be connected to a bond pad. Each box marked "active clamp" represents a circuit that conducts positive current between V_{DD} and V_{SS} only when a fast transient is detected. Positive current between V_{SS} and V_{DD} is carried by the many parasitic N-well to P-substrate diodes throughout the chip, or by a diode the designer places explicitly. The original active clamp circuit is illustrated in Fig. 9.2 [4]. Subsequent versions of this circuit are more compact and/or provide better
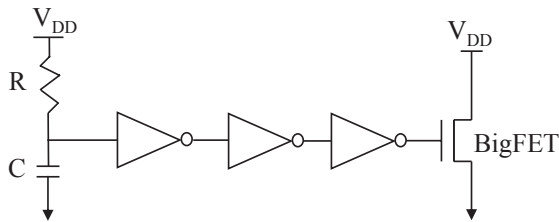

Figure 9.2. Active clamp schematic. Keeping in mind that the IC is unpowered during an ESD event, this circuit turns on when a fast transient (i.e., ESD event) occurs on the V_{DD} line. The RC timer circuit holds the input of the left-most inverter low for an interval on the order of the RC product; consequently, the gate of the BigFET is held at V_{DD}. Once the RC timer charges up to the inverter switching threshold, the gate of the BigFET will be pulled to ground following three inverter delays. The RC time constant is chosen to be on the order of 1-μs, to ensure that the BigFET stays on for the full duration of an HBM event. To ensure that the BigFET does not get driven into breakdown, it must be a very wide device capable of handling large currents while biased in the saturation region of operation. Channel widths as large as 8000 μm are not unusual [8].

latchup immunity [5–7]. It is noted that the BigFET within the active clamp can be implemented as several, moderately sized, nonadjacent, parallel-connected devices; one of these medium-sized FETs would be placed at each I/O pad. Such a distributed network can provide an extremely high ESD protection level [8].

A bi-directional protection circuit between the I/O pad and V_{SS} may be substituted for the dual-diode circuits illustrated in Fig. 9.1. A bi-directional circuit would conduct current of both polarities between its two terminals, although it must remain off when the I/O pad voltage has a value associated with normal operation. Examples of bi-directional protection circuits will be shown later. These tend not to be easily portable between foundries, but are of interest because they potentially provide better voltage clamping for the case of positive stress between the I/O pad and V_{SS} than does the dual-diode circuit [9]. Voltage clamping may be a concern if a thin gate oxide (i.e., MOS FET gate terminal) is connected to the pad. Otherwise, the dual-diode circuit is preferred for RF applications because it has the highest current carrying capability per unit capacitance of all commonly used ESD protection circuits [10].

3. ESD Protection Circuits for RF I/Os

ESD protection circuits load the I/O pads at which they are placed; that is, they introduce a shunt impedence. With this in mind, one may define a figure-of-merit (FOM) for ESD protection circuits that will be used at I/O pads, $FOM = V_{HBM} \cdot Z_{ESD}$ [11], where V_{HBM} is the circuit's failure voltage for Human Body Model stress and Z_{ESD} is its off-state impedance. The off-state impedance of a protection circuit is capacitive, so the FOM may be rewritten as $FOM = \frac{V_{HBM}}{\omega \cdot C_{ESD}}$, where ω is the operating frequency in rad/s. Note that other measures of the circuit's ESD protection level, such as its failure current I_f, may be substituted for V_{HBM}.

A sufficiently high FOM enables the "plug and play" design methodology. Plug and play refers to the practice of placing an ESD protection circuit at the pad after the rest of the design is complete [12]; such practice is reasonable if the last minute inclusion of the protection circuit has an insignificant effect on circuit performance. Note that the FOM is a decreasing function of circuit operating frequency, and we've found that plug and play is extremely difficult at 10 GHz.

A high FOM facilitates co-design. Co-design refers to the practice of designing the RF circuitry and the protection circuit concurrently [13]. Although placing a parasitic load at the pad will have some deleterious effect on performance, co-design enables one to optimize the design such that both reliability and performance are acceptable.

For the case of narrowband I/Os, the cancellation technique—a resonant circuit technique—may be used to relax the FOM requirement for the protection circuit. The T-coil protection circuit is another inductor-based technique for providing ESD protection at RF-inputs, in this case broadband inputs. These protection circuits will be described in Section 4.

3.1 Design for High FOM

The FOM is affected by the choice of protection device, its layout and the interconnect routing. As noted previously, the dual-diode protection circuit has the highest FOM among the protection circuits commonly used in CMOS technology. The dual-diode circuit consists of a "top" diode connected between the I/O pad and the local V_{DD}, and a "bottom" diode connected between the local V_{SS} and the I/O pad. There are several different ways to construct the bottom diode—a substrate diode and an N-well diode are illustrated in Fig. 9.3. In one
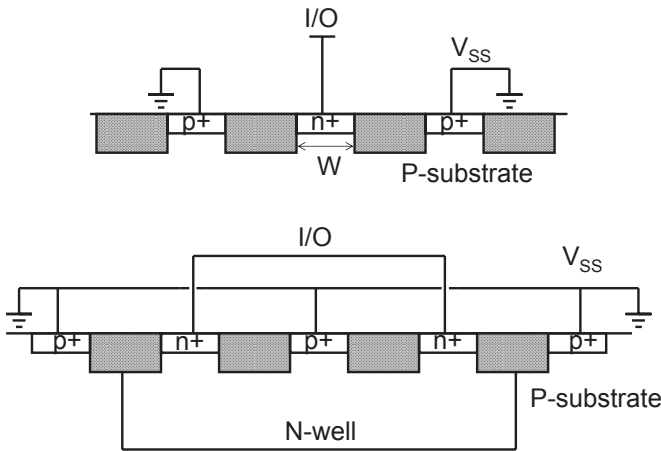
Figure 9.3. A substrate diode is illustrated on the top, and an N-well "bottom" diode is illustrated on the bottom. The substrate diode has a 30% larger FOM. Single stripe designs are illustrated here (for example, there is one N+ stripe shown for the substrate diode) but, in practice, multiple stripes are used to achieve the desired protection level.

experimental study of 0.18-μm RF-CMOS technology [14], it was demonstrated that the substrate diode has a significantly higher FOM than the N-well bottom diode; using a modified FOM of I_f/C_{ESD}, the FOM of the substrate diode was reported to be 29.3 mA/fF, while that of the N-well diode was 20.4 mA/fF.

Electrostatics dictates that current density should be largest along the PN junction perimeter; this is confirmed by the data of Fig. 9.4 which demonstrates that the ESD protection level provided by substrate diodes increases less than linearly with the dimension W shown in Fig. 9.3. As a result, the FOM is a decreasing function of W and minimum (or near minimum) width stripes are recommended. The perimeter effect also explains why the N-well bottom diode has a lower FOM than does the substrate diode. The N-well diode consists of two PN junctions connected in parallel; there is a P^+-diffusion/N-well junction and a P-substrate/N-well junction. The latter junction has a relatively large ratio of area to perimeter; capacitance increases linearly with
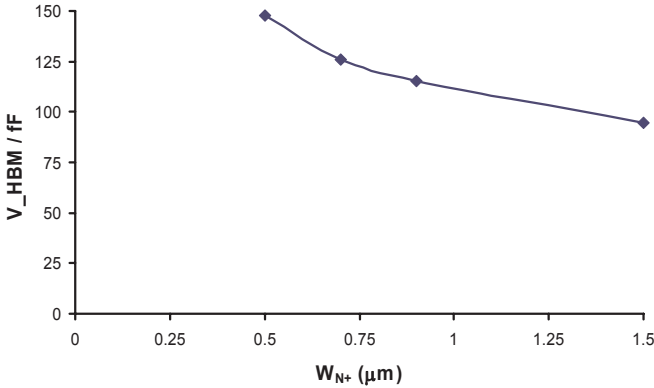
Figure 9.4. Substrate diodes. Protection level per unit capacitance vs. width of N^+ stripe, i.e., the width of the PN junction. The protection level does not scale linearly. Data was provided by E. Worley [15].

area while protection level does not. The well-to-substrate PN junction with its low FOM reduces the overall FOM of the N-well bottom diode. (Note that the N-well top diode has a large FOM [14].)

The interconnects should run in a metal layer that has been chosen to minimize C_{ESD}, i.e., maximize FOM. As illustrated in Fig. 9.5(a), the connection between the diode and the I/O pad can be made using low-level metal. This configuration provides maximum capacitance between the signal line and the grounded substrate. Fig. 9.5(b) illustrates the use of top-level metal to make the connection between the diode and the I/O pad; this minimizes the capacitance between the signal line and ground. In one case study, we found that the routing configuration of Fig. 9.5(b) yields 57% less wire capacitance than that of Fig. 9.5(a). Finally, the connection between the diode and V_{DD} can also be made using top-level metal, as shown in Fig. 9.5(c). Routing V_{DD} on top-level metal does not lower the capacitance relative to that provided by the routing shown in Fig. 9.5(b), since both the V_{DD} line and the substrate are connected to AC grounds. In fact, the configuration of Fig. 9.5(c) has higher capacitance than the one of Fig. 9.5(b) due to the capacitance between the two sets of via stacks. Therefore, the wiring configuration of Fig. 9.5(c) is not recommended.

3.2 Plug and Play

Plug and play is an especially convenient protection strategy for wideband amplifiers. We have demonstrated this for the case of an ultra-wideband (UWB) low noise amplifier (LNA) [16]. Excellent LNAs for UWB receivers can be implemented in SiGe BiCMOS technology. The schematics of common-emitter (CE) and common base (CB) UWB LNAs are shown in Fig. 9.6. Both
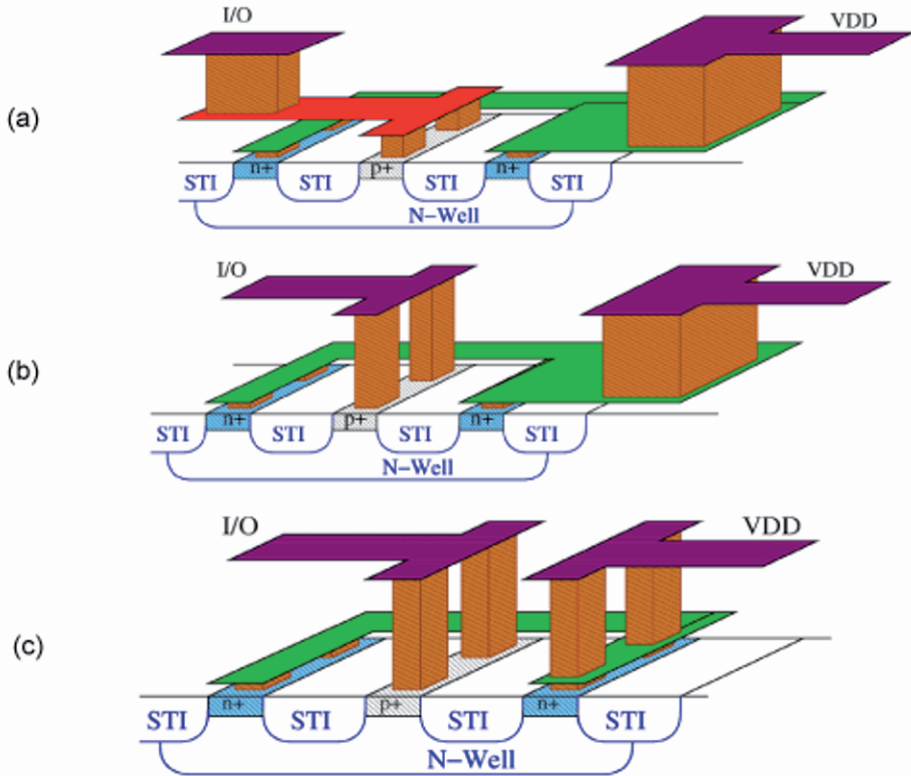
Figure 9.5.   Different wiring configurations for an N-well top diode. (a) Interconnects run on low-level metal. (b) I/O-side interconnect runs on top-level metal. (c) Interconnects run on top-level metal. This configuration is not recommended due to the large capacitance between the two sets of via stacks.

amplifiers can be input matched to 50 Ω over the UWB band (3.1 to 10.6-GHz). For the CE amp, the relevant design equations are $g_m = \frac{f_T}{50 \times 3.1 \times 10^9} \Omega^{-1}$ and $C_{in} = \frac{1}{50 \times 2\pi \times 10.6 \times 10^9} F$ [17], where f_T is the transition frequency of the transistor and C_{in} is the sum of C_{ESD} and the bondpad capacitance. C_{in} must be limited to 300 fF in order for the amplifier to function well up to 10.6-GHz. For the CB amp, the design equation is $g_m = \frac{1}{50} \Omega^{-1}$, and the budget for C_{ESD} depends on the desired value of S_{11}. The total capacitance at the input pad (C_{in}) is the sum of the bondpad capacitance, the current-source capacitance, the emitter capacitance and C_{ESD}. S_{11} is maximum (i.e., worst) at the upper cut-off frequency, 10.6 GHz. The S_{11} and C_{in} are related by $S_{11} = 20 \log \left(\frac{50 \times C_{in} \times \pi \times 10.6 \times 10^9}{\sqrt{1 + (50 \times C_{in} \cdot 10.6 \times 10^9)^2}} \right)$. Typically, S_{11} is desired to be -10 dB or less, in which case C_{in} is limited to 200 fF. Due to the CE amp's relatively
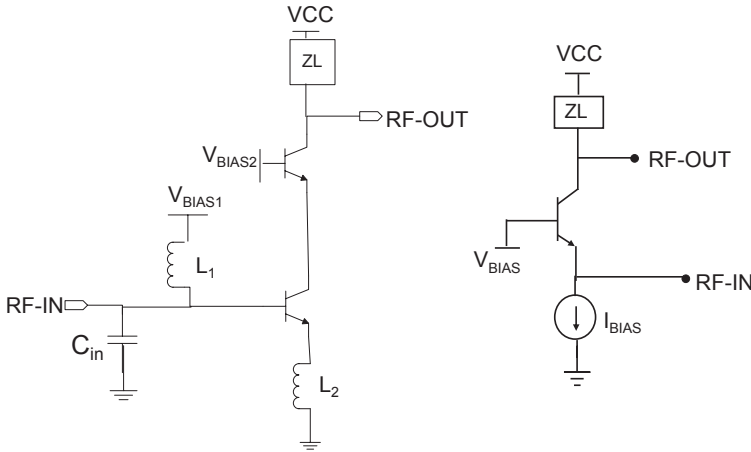
Figure 9.6. The schematic of a wideband CE amp is illustrated on the left [17] and a CB amp is shown on the right. ZL represents the load device, which is usually a series R-L circuit.

Table 9.1. Measurement results for 3 UWB LNAs. The CB designs are from [16]; the CE work is from [17]. The one-inductor CB amp has an R-L load. The zero-inductor CB LNA has a resistive load. The CE amplifier inherently has better performance but consumes considerably more power and area; the area penalty is due to its multiple inductors. CB amps were ESD protected.

	One-inductor CB	Zero-inductor CB	CE cascode
S21	17.0 dB @ 2GHz 19.1 dB at 7GHz	16.1 dB	21 dB
-3dB BW	DC − 10 GHz	DC − 17 GHz	2 − 10 GHz
Noise Figure	4.7dB @ 10 GHz	5.65dB @ 10GHz	4.5dB @ 10GHz
S11	< -10dB	< -10dB	< -10dB
Linearity	-1dbCP = -17.1dBm	-1dBCP = -16.8dBm	IIP3 = 0dBm
VCC	2.7 V	2.7 V	2.7 V
Power dissipation	3.65 mW	3.65 mW	27 mW
ESD Protection	> 1.5 kV	> 1.5 kV	-
Technology	0.18μm SiGe BiCMOS	0.18μm SiGe BiCMOS	0.18μm SiGe BiCMOS

large budget for input capacitance, it is particularly well suited to the plug and play approach for ESD protection. However, plug and play may also be used to protect the CB amp, provided one has available a protection circuit with very high FOM. We have designed and tested an ESD-protected, low-power, UWB CB amplifier; ESD protection was provided by small, robust diodes [15]. Measurement results are summarized in Table 9.1; the results prove that ESD protection does not have to compromise RF performance.

3.3 Co-design

As a first example of co-design, consider the design of a common-source, narrowband LNA; its schematic is shown in Fig. 9.7. Inductive source
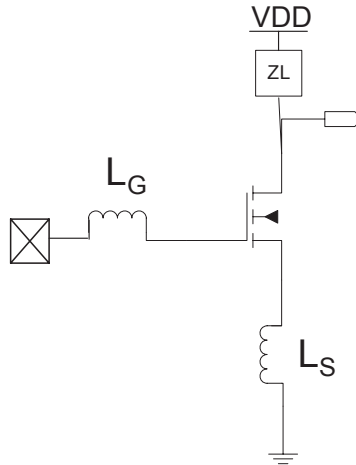

Figure 9.7. Schematic of a common-source, narrowband LNA. Biasing circuitry is not shown.

degeneration is used to obtain the desired input impedance at the circuit operating frequency [18]. For the case of no capacitance at the input pad, i.e., $C_{in} = C_{\mathrm{ESD}} + C_{\mathrm{pad}} = 0$, the relevant design equations are $\frac{L_s g_m}{C_{gs}} = 50$ and $\frac{1}{2\pi}\sqrt{\frac{1}{C_{gs}(L_S+L_G)}} = f_{\mathrm{RF}}$. The input LC circuit also boosts the circuit gain; the amplifier transconductance is given by $G_m = g_m \cdot Q$, where $Q = \frac{1}{50} \cdot \sqrt{\frac{L_S+L_G}{C_{gs}}}$. Even if $C_{in} > 0$, a 50-Ω input match can be obtained by adjusting the values of L_S and L_G, provided C_{in} is not too large. However, in this case, $G_m < g_m \cdot Q$. Clearly, when ESD protection is used, the designer is forced to make trade-offs between gain and input matching. By selecting the ESD protection circuit, and thus C_{ESD}, early in the design cycle, one can then size the input transistor and the inductors, L_S and L_G, so that both S_{11} (input match) and gain (S_{21}) have acceptable values.

Co-design was used for the dual-band, 802.11a/g LNA shown in Fig. 9.8. The input matching network should provide matching and gain at two distinct frequencies; the LLC network shown in Fig. 9.8 was the simplest passive network found to meet these requirements [20]. For the case of no ESD protection, one may derive analytic design equations for L_{M1}, L_{M2}, C_M, L_S, C_{gs}, and g_m [20]. With the inclusion of C_{ESD}, it becomes more practical to use an optimizer to select the component values. Note that small, secondary protection
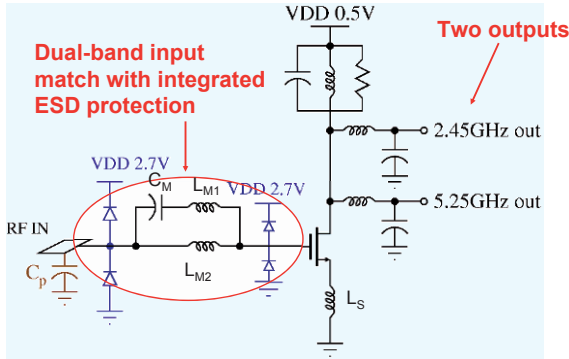
Figure 9.8.   Schematic of a dual-band LNA for 802.11a/g WLAN receivers [19].

diodes were placed at the gate of the input transistor to reduce the likelihood of gate oxide breakdown during Charged Device Model (CDM) ESD events.

The circuit of Fig. 9.8 uses two series-LC loads which are tuned to the two operating frequencies (2.45 and 5.25 GHz). At resonance, these loads minimize the voltage swing at the drain, facilitating highly linear operation at a very low supply voltage. This circuit was fabricated in 0.18-μm BiCMOS technology and was operated using a 0.5-V power supply [19]. Measurement results are summarized in Table 9.2. Overall performance metrics at the upper operating frequency were not as good as expected; this is attributed to mistuning of the upper input match (minimum S_{11} was obtained at 6-GHz rather than 5.25-GHz). However, a very high level of ESD protection was obtained and dual-band operation was demonstrated, thus highlighting the value of co-design.

4.   Inductor-based Protection Circuits

On-chip inductors are area-intensive and, for this reason, forbidden in many cost-sensitive applications. However, on-chip inductors are often used in high-performance RFICs. In such applications, inductor-based ESD protection circuits can be used to achieve an outstanding ESD protection level with virtually no RF performance penalty.

4.1   Cancellation Technique

Two resonant ESD protection circuits are illustrated in Fig. 9.9. The LC resonator protection circuit of Fig. 9.9(a) was proposed by Leroux and Steyaert [21]. It is tuned to the RF operating frequency; at resonance, its impedance is very large (ideally, infinite) and thus its effect on input match is negligible. The inductor is intended to carry the ESD current. The cancellation circuit of Fig. 9.9(b) was proposed in [22]. It contains a traditional ESD protection

Table 9.2.   Dual-band LNA measurement results. Poor performance in the upper frequency band is attributed to mistuning of the input match. Nevertheless, highly linear operation at $V_{\mathrm{DD}} = 0.5V$ has been demonstrated, along with excellent ESD reliability.

	2.45-GHz Operation		5.25-GHz Operation	
	Simulated	**Measured**	**Simulated**	**Measured**
Gain	14.9 dB	13.9 dB	13.1 dB	8.7 dB
NF	2.84 dB	4.98 dB	5.09 dB	6.58 dB
S_{11}	-14.4 dB	-14.4 dB	-19.7 dB	-12.4 dB
IIP_3		+2.87 dBm		+3.23 dBm
-1dBCP	-6 dBm	-6.93 dBm	-2 dBm	-2.26 dBm
V_{HBM}	> 9 kV			
V_{DD}/P_d	0.5 V / 2.5 mW			

(a)   (b)
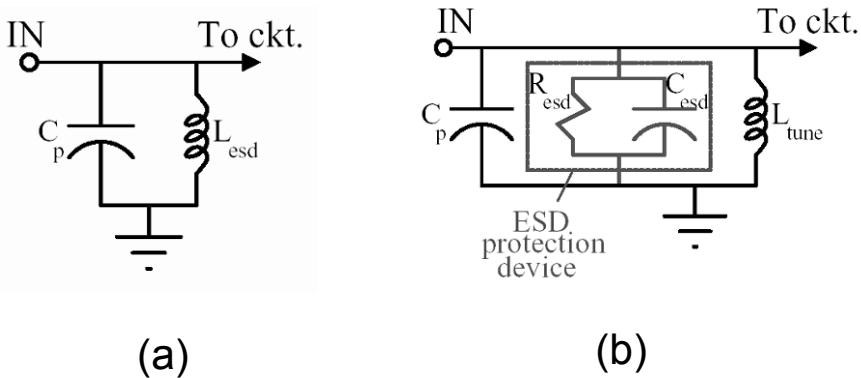
Figure 9.9.   (a) LC resonator protection circuit. (b) Cancellation circuit. R_{esd} and C_{esd} model an ESD protection circuit (e.g., dual-diodes or GGNMOS) in its off-state.

circuit (e.g., dual-diode circuit or grounded-gate nFET) plus a parallel inductor that resonates with C_{ESD} and the bondpad capacitance at the circuit operating frequency. The simulated response of both circuits to a 500-V CDM discharge is shown in Fig. 9.10. The discharge excites high amplitude oscillations at the
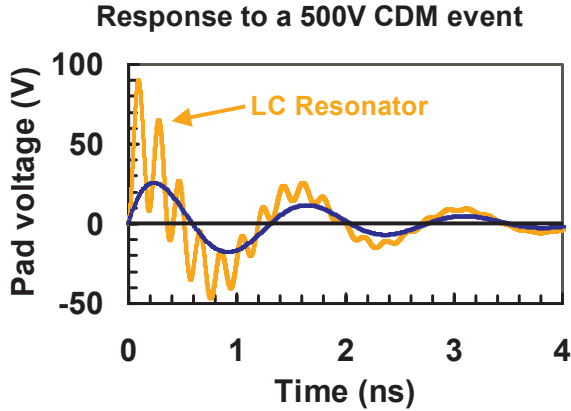
Response to a 500V CDM event

Figure 9.10. The simulated response of the two circuits in Fig. 9.9 to a 500 V CDM discharge.

tank resonant frequency in the LC resonator protection circuit. This is not the case for the cancellation circuit; the low on-resistance of the ESD protection circuit damps the oscillations. Only the cancellation circuit can protect against CDM-type ESD stress.

Under normal operating conditions, the total impedance of the cancellation circuit plus bondpad is given by $Z = \frac{Q_L \cdot Q_C}{Q_L + Q_C} \cdot \frac{1}{\omega C}$, where C is the sum of C_{ESD} and the bondpad capacitance, Q_L is the quality factor of the inductor and Q_C is the quality factor of the capacitor. The FOM of the cancellation circuit is given by $FOM = \frac{Q_L \cdot Q_C}{Q_L + Q_C} \cdot \frac{V_{\text{HBM}}}{C_{\text{ESD}}}$. The cancellation's circuit FOM is larger than that of a "regular" ESD protection circuit by a factor equal to $\frac{Q_L \cdot Q_C}{Q_L + Q_C}$; typically, this factor is about 10. Ideally, the FOM will be limited only by the Q of the on-chip inductor; this means that one must choose high-Q ESD protection devices.

4.2   ESD Protection Devices for the Cancellation Circuit

ESD diodes from a 0.18-μm CMOS technology were characterized in [14]; with one exception, all had acceptably high Q values ranging from 17 to 20 at 5 GHz. The N-well bottom diode, however, only had a Q value of 10.

In another study, the Q values of open-base SiGe HBTs and SiGe dual-diode circuits were simulated [23]. As shown in Fig. 9.11, at a dc bias of $V_{\text{DD}}/2$ and a freqency of 5 GHz, the quality factor of the HBT protection device is about 15 while that of the dual-diode circuit is about 45. The higher Q of the SiGe diodes is attributed to the fact that their p-regions are made exclusively of extrinsic base material; the base-emitter capacitor is thus in series with a smaller resistor than is the base-emitter capacitor of the HBT.
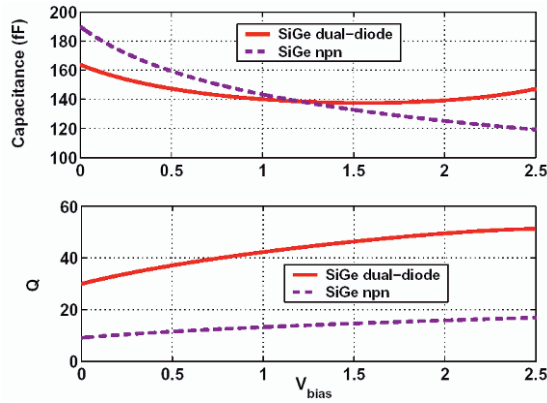
Figure 9.11. Simulated (off-state) impedance at 5 GHz of dual-diode and HBT protection circuits as a function of the pad bias ($V_{DD} = 2.5V$). Devices were sized to provide equal protection levels. The dual-diode circuit has much higher quality factor.

The grounded-gate nFET ("GGNMOS")—an nFET with its source, gate and body tied to ground and its drain tied to the I/O pad—is one of the most frequently used, bi-directional ESD protection devices [2]; this device is normally off, but it can be triggered into breakdown (snapback) by an overvoltage at the drain terminal. GGNMOS are laid out using multiple gate fingers; a gate-coupling circuit ("GCNMOS") may be used to promote uniform turn-on of all the fingers [24]. However, the GCNMOS has very low Q—less than 5 in one experiment [23]—because the gate resistor is in series with the gate-drain capacitance. Uniform turn-on may also be achieved by using a floating body device ("FB-GGNMOS") [25], e.g., a grounded-gate nFET inside an isolated P-well. We observed that deep trench-isolated FB-GGNMOS have 30% higher Q than do conventional, junction-isolated FB-GGNMOS (see Fig. 9.12).

4.3   Cancellation Circuit Optimization

Six versions of a CMOS LNA were designed, each with a different input protection circuit, so that we could quantify the benefits of the cancellation circuit and determine how to best do the implementation [14]. The LNAs were targeted for use in an 802.11a WLAN receiver. They were fabricated in a 0.18-μm RF-CMOS technology. The LNA schematic is shown in Fig. 9.13. The 6 variants are shown in Fig. 9.14.

LNA #1 (Fig. 9.14a) is virtually unprotected. Very small diodes are placed near the gate of the input transistor to provide some protection against CDM stress, and ensure that all the gate oxides were not destroyed during routine handling and before we got to test the parts. These small diodes are present in all the designs. Because the input is not loaded with a large ESD protection
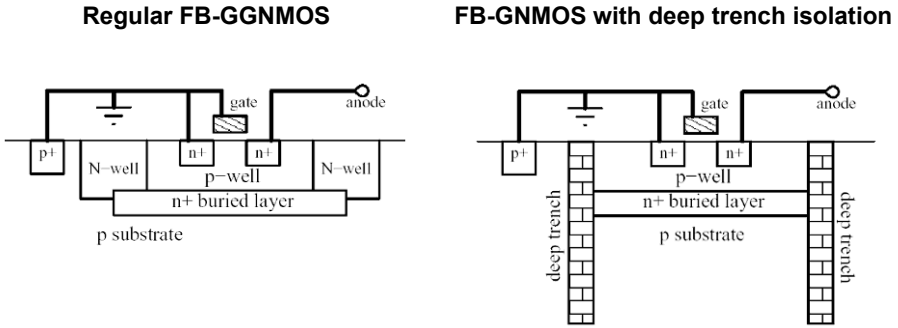
Regular FB-GGNMOS   **FB-GNMOS with deep trench isolation**

Figure 9.12. Regular and deep trench isolated, floating-body grounded-gate nFETs. Quality factor was measured at 5 GHz. The regular device has $Q = 13.3$ and the device with deep trench isolation has $Q = 17.5$.
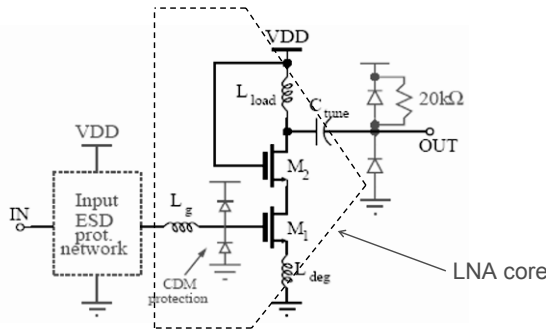

Figure 9.13. Schematic of the LNA test circuit used to evaluate various ESD protection circuits. DC bias circuits are not shown. 0.18-μm RF-CMOS technology. $V_{DD} = 1.8$ V, $P_d = 6$ mW, and $f = 5.25$ GHz.

circuit, we expect LNA #1 to have the best performance. A rail clamp (V_{DD}-to-V_{SS} ESD protection circuit) is present in this LNA and all the others, but is not shown in the schematic.

The input pad for LNA #2 (Fig. 9.14b) is protected with a conventional dual-diode circuit. The diodes are sized to provide at least 4-kV HBM protection. They degrade both the input match and the gain, even after the other circuit components are resized. The output pad is also protected, but its capacitance is absorbed in the output matching network.

LNA #3 (Fig. 9.14c) employs the basic cancellation technique. Note that a dc blocking capacitor is placed in series with the cancellation inductor so as to not short out the dc bias circuit for the input transistor. The dc bias applied to the ESD diodes is the same as that for the amplifier. As a result, the dual
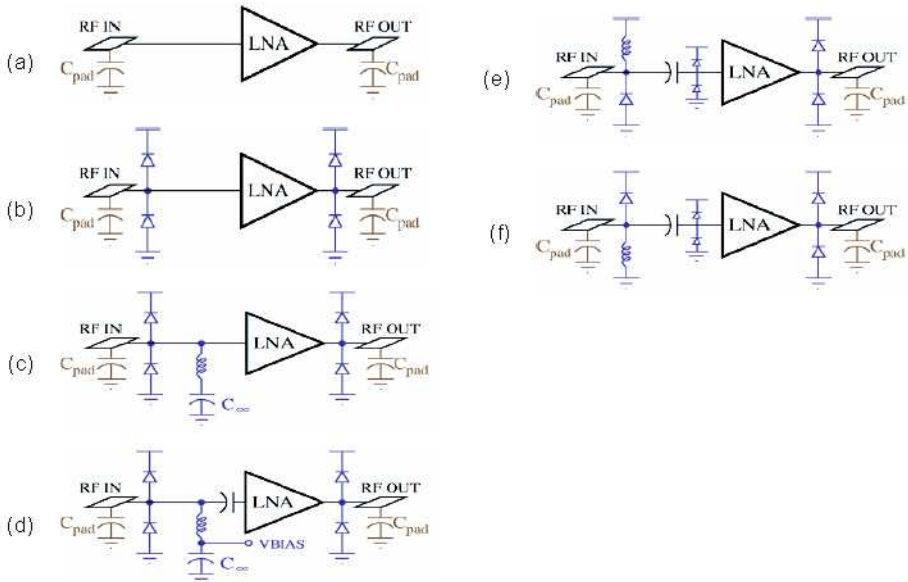
Figure 9.14. (a) Unprotected LNA – #1.(b) Conventional ESD protection – #2.(c) Basic cancellation circuit – #3.(d) Cancellation circuit with optimum dc bias – #4.(e) One-sided, bottom diode protection circuit – #5.(f) One-sided, top diode protection circuit – #6.

diode circuit may not be sitting at its minimum C or maximum Q bias point (see Fig. 9.11 for illustration of these bias dependencies).

In LNA #4 (Fig. 9.14d), a series dc blocking capacitor is included so that the ESD diodes and the input transistor may be biased separately. It was not known whether this circuit's performance would exceed that of LNA #3 because the series MIM capacitor is non-ideal: its bottom plate is capacitively coupled to the substrate (ground); also, the capacitor forms a voltage divider with the LNA.

LNA #5 (Fig. 9.14e) uses the "one-sided, bottom diode cancellation circuit." By replacing the top diode with an inductor, we reduce the total capacitance at the pad and increase the protection circuit FOM. Because the inductor can not protect against fast transients (i.e., CDM stress), the inclusion of the CDM diodes was critical here. Note that this circuit also requires the use of a series dc blocking capacitor.

LNA #6 (Fig. 9.14f) uses the "one-sided, top diode cancellation circuit." In this technology, the top diode has lower C and higher Q than does the bottom diode. Therefore, we anticipated that LNA #6 would have better RF performance than LNA #5.

The measurement results are summarized in Table 9.3. The RF metrics are reported at 5.5-GHz as that was the peak gain frequency. Human Body Model

Table 9.3. Measurement results for the 6 LNAs shown in Fig. 9.14. RF metrics were obtained at 5.5 GHz.

LNA	1	2	3	4	5	6
GAIN(dB)	**12.43**	11.73	12.20	11.94	12.20	**12.41**
NF(dB)	**2.94**	3.65	3.33	3.52	3.31	**3.12**
-1dBCP(dBm)	**-15.0**	-13.4	-14.5	-14.0	-14.3	**-14.6**
IIP_3(dBm)	**-4.16**	-2.90	-3.69	-3.35	-3.58	**-3.72**
S_{11}(dB)	**-16.7**	-8.8	-13.7	-11.8	-13.4	**-14.9**
S_{22}(dB)	**-15.7**	-16.9	-16.5	-16.6	-17.0	**-18.1**
V_{HBM} (kV)	**0.35**	7.6	7.5	7.3	7.1	**7.5**

protection levels were projected on the basis of 100-ns TLP (transmission line pulse) testing, using the formula $V_{HBM} = 1500 \cdot I_{t2}$, where I_{t2} is the second breakdown, i.e., failure, current [26]. All of the circuits with ESD protection are very robust, able to handle more than 7 kV of ESD stress. The unprotected LNA fails at a 350-V level, which is unacceptable; 2 kV is widely accepted as the minimum requirement for HBM protection. The unprotected LNA, however, has the best RF performance measured in terms of gain, noise figure or input match. Only the LNA with conventional ESD protection has unacceptable RF performance; its S_{11} is greater than -10 dB. The cancellation circuits provide significantly better RF performance than does the conventional ESD protection circuit. Despite non-optimum biasing of its ESD diodes, LNA #3 had better RF performance than that of LNA #4. As expected, LNA #6 outperformed LNA #5. Note that the RF performance of LNA #6 is very close to that of the unprotected LNA #1. The only significant difference is the ESD protection level. This study verified that the cancellation circuit provides better RF performance than do conventional ESD protection circuits while providing equally good ESD reliability.

Worst-case ESD protection levels were given in Table 9.3. Table 9.4 provides greater detail; it lists the failure current for each pin combination. In the ESD-protected LNAs, the weakest ESD paths are those containing inductors (although these paths are still very robust, providing over 7-kV HBM protection). Visual inspection revealed that the vias to the inductor underpass were damaged, presumably due to the locally high current density. Therefore, if one wishes to use a one-sided cancellation circuit, one should be careful to use a large number of parallel vias to the underpass.

Table 9.4. Failure current (Amps) for each LNA and for each pin combination. I_{fail} is obtained using 100ns TLP testing. HBM protection level is projected as $1500*I_{fail}$.

LNA	VSS→ IN	IN→ VSS	VDD→ IN	IN→ VDD	IN→ OUT	VDD→ VSS
Unprotected (1)	0.53	0.26	0.54	0.36	0.22	>7.0
Dual Diode (2)	6.7	5.1	>7.0	5.1	5.1	>7.0
Tuned (3)	6.5	5.1	>7.0	4.9	5.0	>7.0
Tuned&Biased (4)	6.1	5.1	>7.0	4.9	5.1	>7.0
1-Sided Bottom (5)	>7.0	4.7	>7.0	5.0	5.3	>7.0
1-Sided Top (6)	5.1	>7.0	4.7	6.4	6.1	>7.0

4.4 T-coil Based ESD Protection Circuit

A T-coil can provide broadband, resistive input matching. Furthermore, it more than doubles the bandwidth relative to that of an RC input, where R is the termination resistor and C is the total capacitance of the ESD protection circuit, bondpad and input device [18, 27]. A schematic of the T-coil based ESD protection circuit [27] is shown in Fig. 9.15. To obtain a resistive input
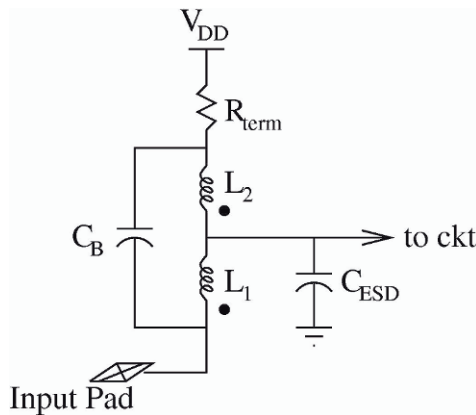
Figure 9.15. T-coil based protection circuit [27]. The circuit input is assumed to be capacitive, e.g., a MOSFET gate terminal. $Z_{in} = R_{term}$ if the values of L_1, L_2, and C_B are chosen correctly.

match, one must select L_1, L_2, and C_B according to the design equations in [27]. In the one study of this protection circuit, a 1-kV HBM ESD protection level was achieved [27]. We note that this is not an adequate protection level, and that a similar protection level with far less area penalty could be obtained by

using a too-small, conventional ESD protection circuit. Nevertheless, the T-coil should still be considered a promising ESD protection technique. Further study is needed to identify the cause of the low protection level reported in [27] and implement a remedy.

5. ESD Testing of RFICs

ESD-induced failures are generally detected by monitoring the leakage current. Leakage testing can not be performed on circuits containing certain variants of the cancellation circuit, specifically, those with an inductor directly connected between the input pad and V_{SS}. Of more general concern is the fact that RF performance metrics might be degraded at stress levels lower than that which causes a significant increase in leakage [28, 29]. To investigate this subject, a combined RF and TLP test system was developed. A TLP (transmission line pulse) tester generates short duration, high current pulses by discharging a charged transmission line into the device under test [30]. ESD can occur between any arbitrary pair of pins so one must be capable of doing TLP testing between any pair of pads. To accomplish this, the TLP return path had to be isolated from the chip ground to which all the RF test equipment is connected. The resulting measurement system is illustrated in Fig. 9.16 [31]. In several test
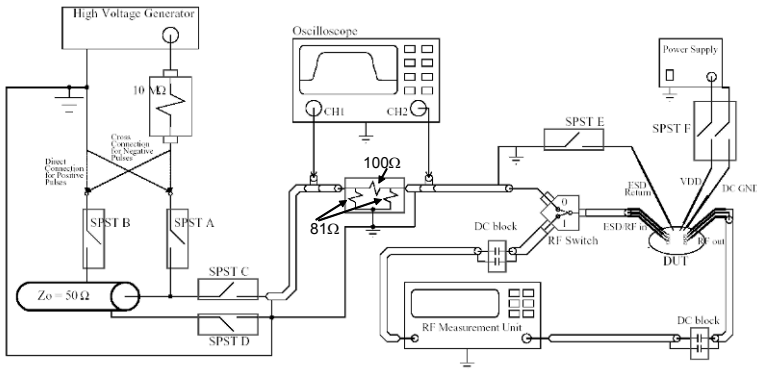
Figure 9.16. TLP/RF measurement system [31]. The RF measurement unit can be any piece of RF test equipment, e.g., a noise figure meter or VNA. Note that the resistor box, i.e., resistive π-network, for the TLP system has $50\,\Omega$ input impedance at both ports. As a result, the pulse does not bounce back and forth between the DUT (device under test) and the resistor box, facilitating the production of square pulses with short risetime. Note that the TLP system ground is isolated from the chip ground. This allows for ESD testing between any arbitrary pair of pads on the chip. During RF testing, only switches F and C are closed (and the RF switch is in the lower position). Switch F is open during TLP testing.

circuits, a change in the RF performance metrics (gain, noise figure) occurred at a current level that was 20% lower than the failure current as determined by

leakage measurement. Therefore, it is recommended that RF performance be monitored during ESD testing.

Acknowledgments

References

[1] C. Duvvury and A. Amerasekera, "ESD: a pervasive reliability concern for IC technologies," *Proc. IEEE*, vol. 81, no. 5, pp. 690–702, 1993.

[2] A. Amerasekera and C. Duvvury, *ESD in Silicon Integrated Circuits*, 2nd ed., John Wiley & Sons, 2002.

[3] J. Miller, "SPICE-based ESD protection design utilizing diodes and active MOSFET rail clamp circuits," *2005 EOS/ESD Symposium Tutorial*, 2005.

[4] R. Merrill and E. Issaq, "ESD Design methodology," *Proc. EOS/ESD Symp.*, pp. 233–237, 1993.

[5] J. Smith and G. Boselli, "A MOSFET power supply clamp with feedback enhanced triggering for ESD protection in advanced CMOS processes," *Proc. EOS/ESD Symp.*, pp. 8–16, 2003.

[6] M. Stockinger, *et al.*, "Boosted and distributed rail clamp networks for ESD protection in advanced CMOS technologies," *Proc. EOS/ESD Symp.*, pp. 17–26, 2003.

[7] J. Li, R. Gauthier, and E. Rosenbaum, "A compact, timed-shutoff, MOSFET-based power clamp for on-chip ESD protection." *Proc. EOS/ESD Symp.*, pp. 273–279, 2004.

[8] C. Torres, J. Miller, M. Stockinger, M. Akers, M. Khazhinsky, and J. Weldon, "Modular, portable and easily simulated ESD protection networks for advanced CMOS technologies," *Proc. EOS/ESD Symp.*, pp. 82–95, 2001.

[9] K. Chatty, *et al.*, "Study of factors limiting ESD diode performance in 90nm CMOS technologies and beyond," *Proc. Int. Rel. Phys. Symp.*, pp. 98–105, 2005.

[10] C. Richier, *et al.*, "Investigation on different ESD protection strategies devoted to 3.3 V RF applications (2 GHz) in a 0.18 μm CMOS process," *Proc. EOS/ESD Symp.*, pp. 251–259, 2000.

[11] S. Hyvonen, S. Joshi, and E. Rosenbaum, "Comprehensive ESD protection for RF inputs," *Proc. EOS/ESD Symp.*, pp. 188–194, 2003.

[12] S. Thijs, *et al.*, "ESD protection for a 5.5 GHz LNA in 90 nm RF CMOS - implementation concepts, constraints and solutions," *Proc. EOS/ESD Symp.*, pp. 40–49, 2004.

[13] V. Vassilev, *et al.*, "Co-design methodology to provide high ESD protection levels in advanced RF circuits," *Proc. EOS/ESD Symp.*, pp. 195–203, 2003.

[14] S. Hyvonen and E. Rosenbaum, "Diode-based tuned ESD protection for 5.25-GHz LNAs," *Proc. EOS/ESD Symp.*, pp. 9–17, 2005.

[15] E. Worley and A. Bakulin, "Optimization of input protection diode for high speed applications," *Proc. EOS/ESD Symp.*, pp. 62–72, 2002.

[16] K. Bhatia, S. Hyvonen, and E. Rosenbaum, "A compact, ESD-protected, SiGe BiCMOS LNA for ultra-wideband applications," manuscript in preparation.

[17] A. Ismail and A. Abidi, "A 3 to 10 GHz LNA using a wideband LC-ladder matching network," *Proc. IEEE Int. Solid-State Circuits Conf.*, pp. 384–385, 2004.

[18] T. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, 2nd ed., Cambridge University Press, 2004.

[19] S. Hyvonen, K. Bhatia, and E. Rosenbaum, "An ESD-protected 2.45/5.25-GHz dual-band CMOS LNA with series LC loads and a 0.5-V supply," *Proc. IEEE RFIC Symp.*, pp. 43–46, 2005.

[20] S. Hyvonen, *Electrostatic Discharge Protection and Measurement Techniques for Radio Frequency Integrated Circuits*, Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2004.

[21] P. Leroux and M. Steyaert, "High-performance 5.2GHz LNA with on-chip inductor to provide ESD protection," *Elec. Lett.*, vol. 37, no. 7, pp. 467–469, 2001.

[22] S. Hyvonen, S. Joshi, and E. Rosenbaum, "A cancellation technique to provide ESD protection for multi-GHz RF inputs," *Electronics Letters*, vol. 39, no. 3, pp. 284–286, 2003.

[23] S. Joshi, S. Hyvonen, and E. Rosenbaum, "High-Q ESD protection devices for use at RF and broadband I/O pins," *IEEE Trans. Elec. Dev.*, vol. 52, no. 7, pp. 1484–1488, 2005.

[24] C. Duvvury and C. Diaz, "Dynamic gate coupling of NMOS for efficient output ESD protection," *Proc. IEEE Int. Rel. Phys. Symp.*, pp. 141–150, 1992.

[25] S. Joshi, P. Juliano, E. Rosenbaum, G. Kaatz, and S. M. Kang, "ESD protection for BiCMOS circuits," *Proc. IEEE Bipolar/BiCMOS Circuits and Technol. Mtg.*, pp. 218–221, 2000.

[26] J. Barth, K. Verhaege, L. Henry, and J. Richner, "TLP calibration, correlation, standards and new techniques," *IEEE Trans. Elec. Pkg. Mfg.*, vol. 24, pp. 99–108, 2001.

[27] S. Galal and B. Razavi, "Broadband ESD protection circuits in CMOS technology," *IEEE J. Solid-State Circuits*, vol. 38, no. 12, pp. 2334–2340, 2003.

[28] S. Voldman, et al., "Test methods, test techniques and failure criteria for evaluation of ESD degradation of analog and radio frequency (RF) technology," *Proc. EOS/ESD Symp.*, pp. 92–100, 2002.

[29] V. Vassilev, G. Groeseneken, S. Jenei, R. Venegas, M. Steyaert, and H. Maes, "Modelling and extraction of RF performance parameters of CMOS electrostatic discharge protection devices," *Proc. EOS/ESD Symp.*, pp. 111–118, 2002.

[30] T. Maloney and N. Khuruna, "Transmission line pulsing technique for circuit modeling of ESD phenomena," *Proc. EOS/ESD Symp.*, pp. 49–54, 1985.

[31] S. Hyvonen, S. Joshi, and E. Rosenbaum, "Combined TLP/RF testing system for detection of ESD failures in RF circuits," *IEEE Transactions on Electronics Packaging Manufacturing*, vol. 28, no. 3, pp. 224–230, 2005.

Chapter 10

SILICON-BASED MILLIMETER-WAVE POWER AMPLIFIERS

Mona Mostafa Hella and Burak Çatli

1. Introduction

Millimeter-wave sensor and communication applications create an increasing demand for high-frequency devices and monolithic integrated circuits. Compared to the radio frequency band which is highly occupied with cellular and WLAN standards, the millimeter-wave band offer the availability of broader frequency ranges, higher gain and smaller antenna dimensions. For sensor applications, the shorter wavelength results in higher resolution and for communication purposes, the atmospheric attenuation at these frequencies decreases interference between the communicating cells [1].

Traditionally, systems operating in the high microwave/mm-wave bands are realized using multiple modules implemented mainly in III-V technologies, increasing the overall cost and complexity. The recent advances in silicon-based technologies have generated devices with transit frequencies (f_T) in the 100-200 GHz range thus extending the reach of these technologies to the mm-wave circuit and system development. Silicon is by far the most widely used semiconductor that provides low cost, established production lines and a high standard of technological expertise. Integration in silicon will make it possible to realize complex mm-wave transceivers with on-chip mixed-signal and digital signal processing, at much lower costs and with high reliability and yield. Additionally, complex analog/RF/mm-wave subsystems can be digitally tuned for optimized performance under different operating conditions. These factors provide a strong motivation for the development of new circuit and architectural techniques that overcome the drawbacks of low active gain devices and low

quality factor passives in standard silicon processes for the implementation of mm-wave systems.

Monolithic integrated millimeter-wave circuits based on CMOS and SiGe technologies will dramatically impact the cost, size, and availability of millimeter-wave applications such as radars and gigabit wireless local area networks. Fully integrated phased array radar transceivers operating at 24GHz as well as front-end components at 60 and 77GHz bands have been recently published in CMOS and SiGe technologies [2]- [7]. This chapter will focus on the challenges and possible design techniques of power amplifiers for millimeter-wave transceivers.

2. Challenges in Microwave/Millimeter-Wave Power Amplifier Design

The design of a fully integrated power amplifier with reasonable output power, efficiency and gain remains a challenge in today's pursuit of system-on-chip (SOC) RF and mm-wave integrated transceivers. Although CMOS and SiGe devices have excelled in small signal (output power < -10dBm) mm-wave circuits [8]-[10], silicon-based power amplifiers have rarely delivered power above 10-20dBm with power added efficiency PAE above 15% at frequencies beyond 10GHz.

The low output power and efficiency in silicon-based mm-wave power amplifiers are due to the limitations of the active devices and the lossy passive network used for impedance matching and wave-shaping. For active devices, the maximum output power of the single device is limited by the value of the breakdown voltage. The breakdown voltage of CMOS and SiGe transistors with $f_{max} > 70$GHz for mm-wave applications is typically less than 2V, while III-V transistors with equivalent performance have a breakdown voltage higher than 8V[11]. In addition, silicon-based transistors have lower gain and thus require higher DC-bias currents which translates into a lower value of optimum impedance ($R_{opt} \approx \frac{V_{max}}{I_{max}}$) in order to deliver the same amount of power as III-V devices. This high bias current constraint the design of on-chip passives because of the metal electro-migration at high current densities. Thus, the width of the metal-lines forming on-chip inductors (RF chokes and output matching elements) has to increase to handle these large currents, thus increasing the parasitic capacitance to ground and lowering the self-resonance frequency. The reduction in R_{opt} has also a practical lower limit determined by the impedance matching to an off-chip 50Ω load. The higher the transformation ratio, the more sensitive it is to component tolerances which affects both the loss and the bandwidth of the matching network.

In addition to the limitations posed by active devices, the low substrate resistivity and the thin metal and dielectric layers in a typical silicon-based process,

degrade the quality factor (Q) of on-chip inductors and capacitors. Moreover, at higher frequencies, skin effect increases the ohmic losses of inductors and transmission lines. This is due to the reduction of effective metal thickness which limits the benefits gained by using thicker metal layers in RF-enhanced processes.
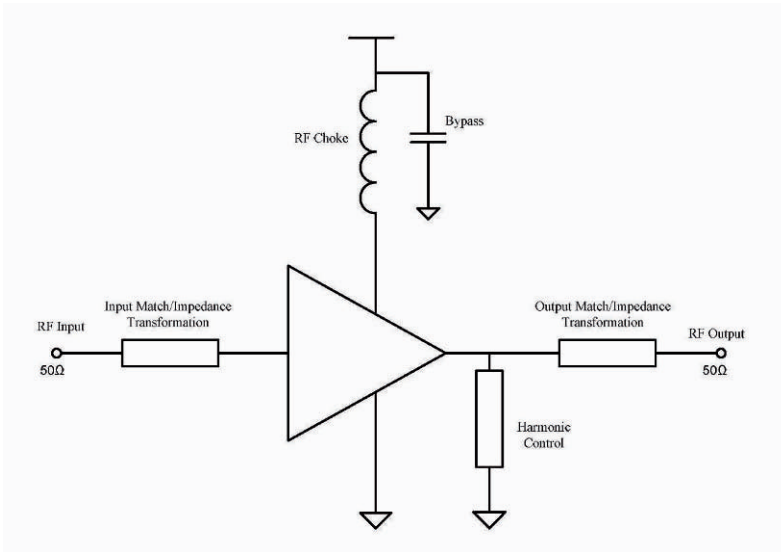

Figure 10.1. Typical power amplifier schematic.

As shown in Figure 10.1, there are several on-chip components in a typical power amplifier. First, an inductor is required in the harmonic control block to achieve the desired harmonic suppression at the output and to improve the drain wave shaping for better power efficiency and linearity by resonating the transistor's drain capacitance. The low Q of this inductor can have a significant effect on the power efficiency of the power amplifier, typically losing 20% to 40% of the power drained from the DC supply in this inductor alone. In order to minimize the loss, the impedance of this inductor should be comparable to the low impedance of the load presented to the transistor and the output impedance of the transistor.

A high impedance inductor is needed as an RF choke to feed the DC current to the transistor and, meanwhile, maintain the low-impedance, low-loss AC ground together with the bypass capacitor. However, due to the series metal resistance, this inductor, if implemented on-chip, suffers from a significant power loss. The bypass capacitor should also be large, which would consume a large area on chip.

In the input and inter-stage matching networks, it is necessary to place an inductor in parallel with the gate to tune out the gate capacitance in order to match the low gate impedance to the 50Ω input or to the output impedance of the driver stage. A low impedance DC block, which will be implemented by a big capacitor, is also necessary in the input matching network.

For mm-wave frequencies, the size of the different inductors in the power amplifier circuit falls in the 10-100pH range. Inductor-lines, microstrip, and coplanar wave-guide structures are typically used for low-inductance values implementation. Extensive electromagnetic simulations are required to model all passives as well as interconnect lines whose inductance values are comparable to those used in the basic amplifier.

3.   Power Amplifier Design Approaches

The choice of the power amplifier architecture is determined by the output power, gain, and PAE requirements as well as the used technology (f_{max} relative to the operating frequency). As the operating frequency increases relative to f_{max} of the transistors, the gain of the single stage decreases. To reach the required output power levels, the number of gain stages increases and the overall power added efficiency (PAE) decreases. Multi-stage amplifiers are also more prone to instability. Harmonic matching based classes, such as class E or class F, do not help in increasing the efficiency in this case as the harmonic content at the drain of the transistor is already low at the higher frequencies close to f_{max}.

We will start with the single-ended amplifier stages and address techniques to increase their output power capability while using low breakdown voltage transistors. We will then move to differential stages and finally to multi-way power combining techniques.

3.1   Single-Ended Amplifier Topologies

Single-ended designs are used to avoid using a balun or a differential antenna. The effect of ground bondwire inductances is one of the main problems with single-ended topologies at millimeter-wave frequencies. For example, a bondwire inductance as low as 50pH at 24GHz adds 7.5Ω reactive impedance to the on-chip ground, which is comparable to the optimum load for a silicon-based common source/emitter stage. This source/emitter degeneration reduces the gain and the power added efficiency (PAE) of the amplifier. In a multi-stage single-ended power amplifier, the effect of ground bounce might cause stability issues as well. One possible solution is to increase the number of ground pads and separate the grounds of the different stages.
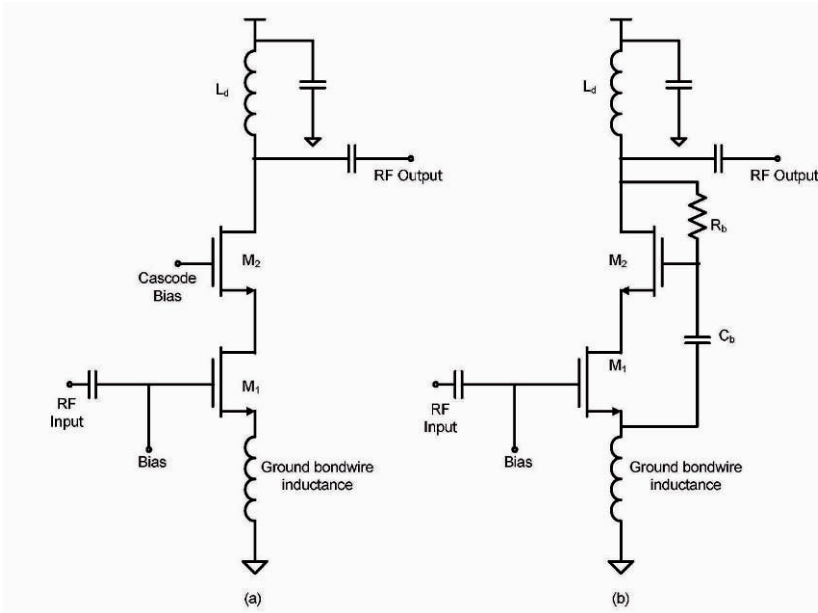
Figure 10.2. (a) Conventional cascode amplifier. (b) Self-biased cascode amplifier[14].

Due to the low breakdown voltage limitation, the cascode configuration using thick-oxide transistors is widely used in CMOS technology[12, 13]. For mm-wave frequencies, the thick-oxide transistor option might not be feasible as its f_{max} is much lower than thin-oxide transistors which decreases the power amplifier gain. However, cascoding can be done using two high f_T transistors. In the traditional cascode structure as the one shown in Figure 10.2(a), the biasing at the gate of the cascode transistor is fixed. Thus, most of the output voltage still appears across the upper transistor. The self-biased cascode technique, shown in Figure 10.2(b) has been proposed by Sowlati et al [14, 15] to extend the output power capability of the output stage by equally distributing the output voltage across the main and the cascode transistors. The gate bias for transistor M_2 is provided by R_b-C_b, which is the same as the DC drain voltage of M_2. The RF swing at the drain is attenuated by the low pass nature of R_b-C_b. The values of R_b and C_b are typically chosen for equal gate drain signal swings on M_1 and M_2. Thus, hot carrier effects and oxide breakdown are avoided.

The concept of RF-driven cascoding delivers the same effect of equal voltage distribution through using a different cascoding approach. For SiGe HBT transistors, in a standard cascode configuration, if the base of the upper transistor (Q_2) is biased at a constant voltage, as shown in Figure 10.3(a), through a filtered regulated current mirror (not shown in the figure), the emitter voltage will remain relatively constant due to the logarithmic dependence of V_{be}

on I_c. Most of the output swing will still appear across the upper transistor, which will cause the upper transistor to suffer breakdown while the bottom transistor (Q_1) will have a relatively small collector-emitter voltage (V_{ce}) swing. Figure 10.3(b) shows the collector and emitter voltage waveforms for the fixed-bias cascode configuration, where the emitter of Q_2 remains almost constant and Q_2 experiences high collector-emitter voltage. By connecting the base of
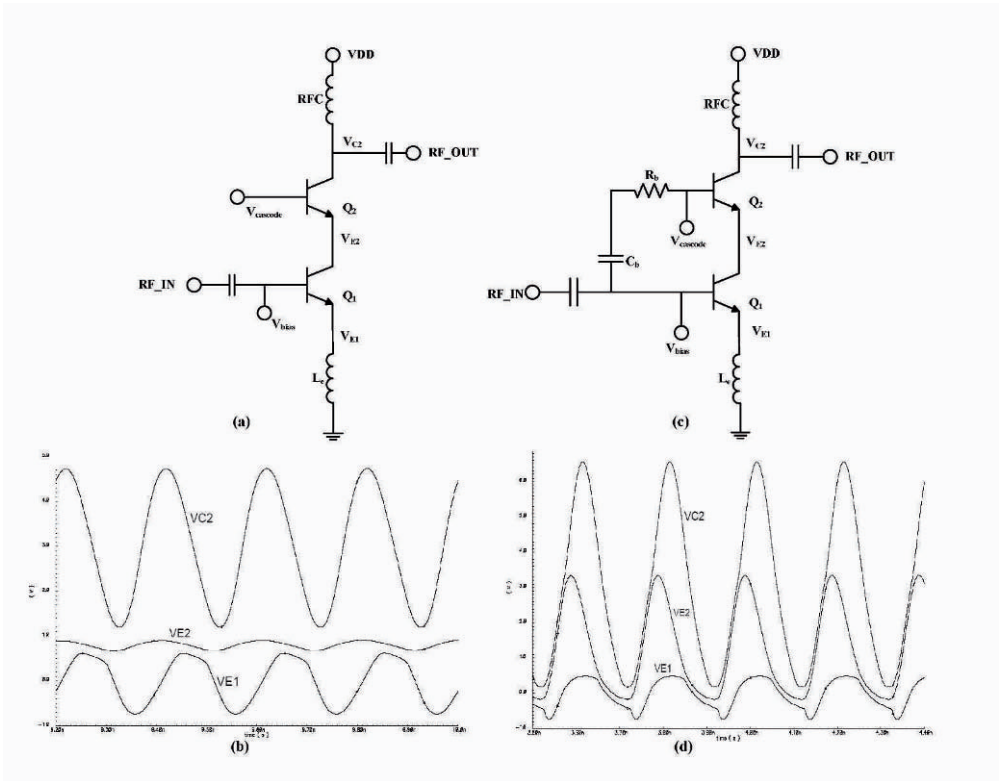
Figure 10.3. (a) Fixed-bias cascode amplifier. (b) Collector and emitter voltage waveforms for fixed-bias cascode amplifier. (c) RF-driven cascode amplifier. (d) Collector and emitter voltage waveforms for the RF-driven cascode amplifier.

the cascode transistor (Q_2) to the RF input through an RC network as shown in Figure 10.3(c), the base will move with the same polarity as the output voltage. The emitter of Q_2 will follow the base simultaneously as an emitter-follower for the base drive, which will effectively divide the output voltage swing between the top (Q_2) and the bottom (Q_1) transistors, thus decreasing the stress on the collector-emitter of Q_2. Ideally, the base of Q_2 should move with half the voltage swing of its collector (V_{out}) for equal voltage division between the upper and lower transistors. It can be observed in Figure 10.3(d) that V_{CE2} is

almost equal to V_{CE1}, effectively distributing the stress on both Q_1 and Q_2. In addition, the low-impedance path at the base of Q_2 increases its breakdown voltage BV_{cer} and extends its output power range [16].

Example: A 24GHz Single-Ended Power Amplifier
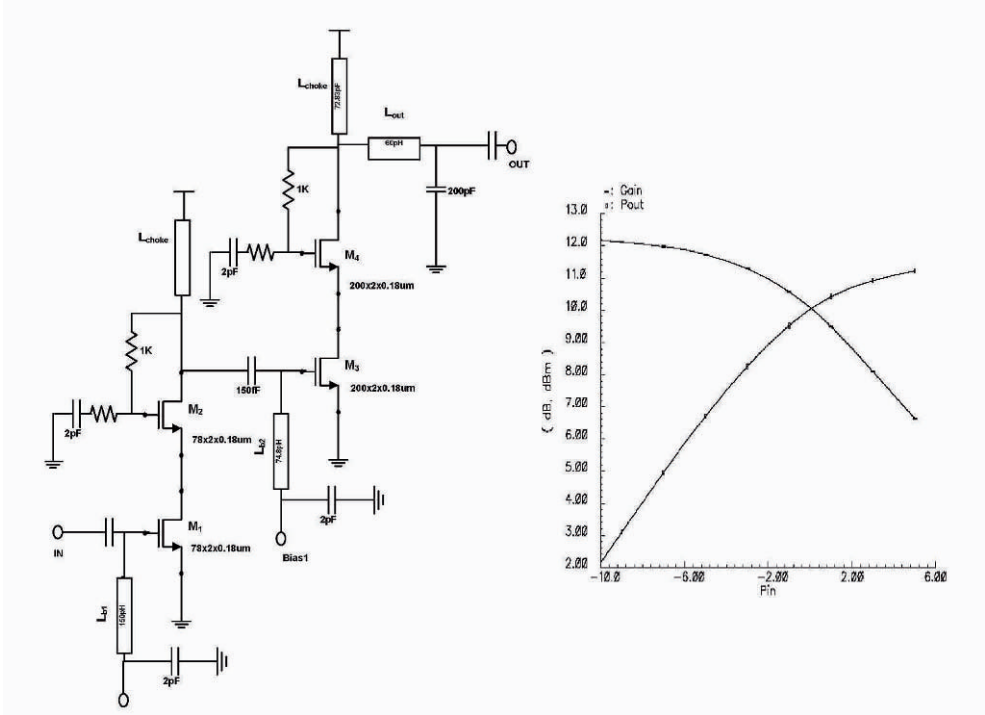
Figure 10.4. (a) Schematic of a two stages single-ended CMOS power amplifier. (b) Simulated output power and power gain of the amplifier at 24GHz.

There are only three mm-wave power amplifiers reported to date in silicon-based processes[11]-[17, 18], with only one as a single-ended amplifier in CMOS technology[17]. For the this design, a 2-stage single-ended class AB power amplifier is implemented using 0.18μm FETs. A cascode configuration is used in each stage to ensure stability and increase breakdown voltage. The self-biased technique [14] is used for the cascode transistors to extend the output power range. To minimize the effect of gate resistance, which can be a limiting factor for f_{max}, the finger width of the unit transistor was chosen to be very small with large number of fingers (2μm×200) and gate contacts at both ends. This also allows substrate contacts to be placed closer to the device, minimizing substrate losses.

For the schematic shown in Figure 10.4(a), input, output, and inter-stage matching are implemented on-chip using inductor-lines formed of the top metal layer over a deep trench to isolate the inductor from the substrate. This technique generates small value inductors with high quality factor. In the original design in [17], the inductor lines are replaced by shielded substrate coplanar wave guide structures. The width of all inductor lines is determined based on the current handling capabilities of these lines. Inductors L_{b1} and L_{b2} tune out the capacitance at the gates of transistors M_1 and M_3. Inductors L_1 and L_2 act as RF chokes, while L_{out} transforms the 50Ω load to the optimum impedance at the drain of M_4.

The amplifier has been simulated while accounting for the effect of parasitics, including ground inductances as shown in Figure 10.4(b). Using 5 ground bonding pads with their typical packaging parasitic inductances, the PA can deliver 11dBm of maximum output power. The output power is estimated to increase to 14dBm with around 6dB increase in gain by decreasing the ground inductance. This shows the enormous effect the ground bondwires have on the gain and output power capability of the power amplifier.

3.2   Cascaded Differential Stages with Transformer-Based Inter-stage Matching and On-Chip Output Balun

The differential push-pull topology can be used to eliminate many problems in single-ended configurations. As shown in Figure 10.5, due to its symmetric configuration, AC grounds are created at both DC supply and on-chip ground nodes. They are inherently low loss and low impedance. Thus, the need for a lossy on-chip RF choke inductor and a large bypass capacitor is avoided. The connection from these ac virtual grounds to the positive supply and ground will carry only current at dc and even harmonics, thus eliminating the loss caused by the RF signal and odd harmonics going through lossy supply lines. Furthermore, this structure desensitizes the operation of the amplifier to external bond wires.

The only limitation is the need for an output balun to transform the differential load into 50Ω impedance. While off-chip baluns are typically used at lower frequency ranges, on-chip L-C baluns are a compact and integrated solution that can be employed in the mm-wave frequency range.

As an example, the power amplifier shown in Figure 10.5 targets the 17GHz band [19]. It employs two on-chip transformers for the input balun and for interstage matching. The push-pull circuit configuration has a 4:1 load-line impedance benefit compared to a single-ended design, which is a main advantage at low supply voltages as discussed in section 2.

The input transformer acts both as a balun and as an input matching network when combined with capacitors C_1 and C_2. The coupling coefficient in the design published in [19] is limited to 0.45 at 17GHz which introduces major losses between stages and reduces the overall output power. This is mainly due
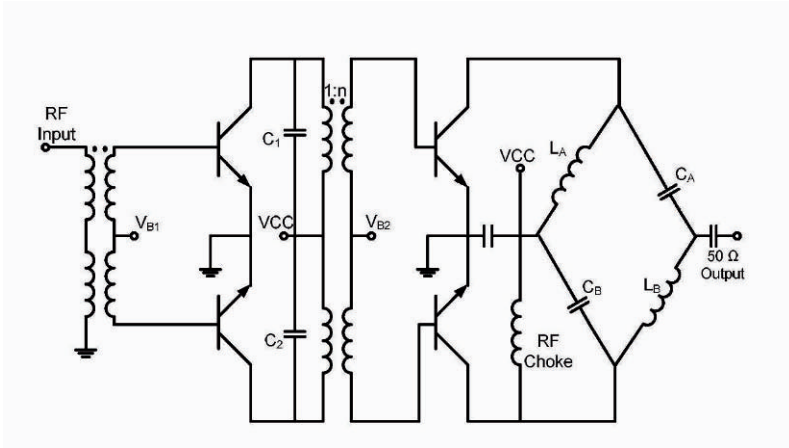
Figure 10.5. A fully integrated differential SiGe PA with transformer-based matching network and on-chip balun [19].

losses between stages and reduces the overall output power. This is mainly due to the limitation on the number of metal layers in the used technology (three metal layers).

A Lattice-type LC-balun [20] is used to provide output matching while converting the differential signal to 50Ω. While this design is based on lumped components, transmission-line implementation has also been reported [21]. The ideal lattice-type balun consists of two inductances $L_A = L_B$ and two capacitors $C_A = C_B$ (Figure 10.5). An RF choke and a DC-block capacitor are added to bias the transistors. If R_{opt} is the optimum balanced impedance at the collectors of the push-pull pair, and R_L is the load impedance, then L and C can be calculated according to:

$$L_A = L_B = \frac{Z_1}{\omega} \tag{10.1}$$

$$C_A = C_B = \frac{1}{\omega Z_1} \tag{10.2}$$

$$Z_1 = \sqrt{R_{opt} R_L} \tag{10.3}$$

where Z_1 is the characteristic impedance of the bridge and $\omega = 2\pi f$ is the frequency of operation. Asymmetry in the balun response due to the effect of bond-wire inductance connected to the output pads and parasitic capacitances

causes the transformed load impedance to appear inductive at one of the balanced output ports and capacitive at the other. Compensating for this effect requires careful matching of the transistors connected to each port of the balun by using unequal values for the passive components in the bridge. Thus, a symmetrical balun design is required for mm-wave power amplifiers to ensure the transistors drive identical loads over a broad frequency range.

Implemented in $0.35\mu m$ SiGe technology, the power amplifier in this example can operate from 8GHz to 17GHz without any external element. It delivers 17dBm of output power using a 2.4V supply with a small signal gain of 15dB.

4.   Power Combining Techniques

When the required output power is larger than what is available from a single transistor or a differential stage, power combiners can be used to sum the power from several transistors or amplifier stages to achieve the desired overall output power over a wide bandwidth. Traditional microwave power combining techniques rely on the properties of wave-guides to make effective multi-way power combiners with minimum losses. In microwave monolithic integrated circuit designs (MMIC), the Wilkinson power combiner is one possible technique that is realized using $\lambda/4$ transmission lines [22]. The realization of effective integrated power combiner is determined by the size and quality factor of the on-chip passive network. Two possible approaches have recently evolved; the doubly-balanced architecture using directional couplers, and the distributed active transformer.

4.1   Doubly-Balanced Architectures

Balanced amplifiers are a classical microwave concept that is typically used to provide a flat gain over wide bandwidth. A fairly flat gain is obtained if the amplifier is designed for less than maximum gain at the expense of poor input and output matching. By using two $90°$ hybrid couplers to cancel out input and output reflections from two identical transistors or amplifier stages, matching can be restored while maintaining a flat gain. The first $90°$ hybrid coupler divides the input signal into two equal amplitude components with a $90°$ phase difference, while the output coupler recombines them. Because of the phasing properties of the hybrid coupler, reflections from the amplifier inputs cancels at the input of the hybrid coupler, resulting in improved impedance match; a similar effect occurs at the output of the balanced amplifier. The doubly balanced amplifier combines two balanced amplifier in a differential configuration. The $180°$ hybrid coupler is used to provide a differential signal for the two balanced amplifiers. The architecture has recently been reported in [23] for WLAN applications targeting the 5GHz frequency band. The output power in this case is quadrupled compared to that of a single amplifier. The topology also allows the
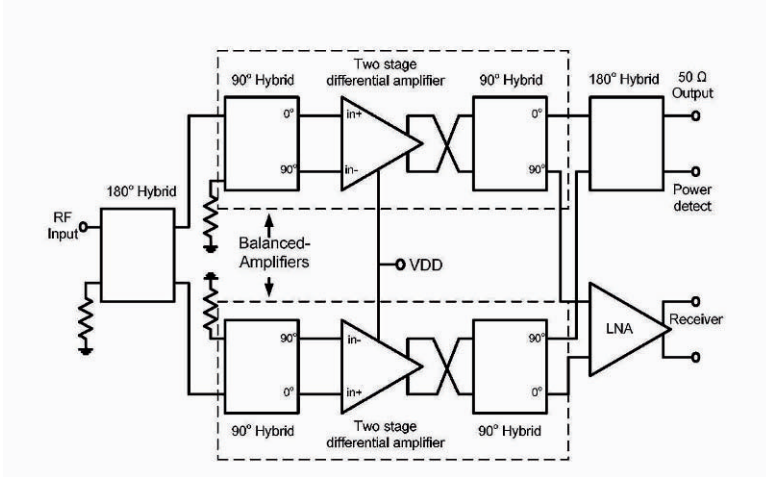
Figure 10.6. Doubly-balanced transceiver front-end topology based on 90° and 180° couplers [23].

implementation of an antenna switch suitable for time division duplex (TDD) systems without any additional components. A fully differential transceiver architecture is maintained using this topology as well as the availability of one port to monitor the output power of the amplifier. The main challenge in this
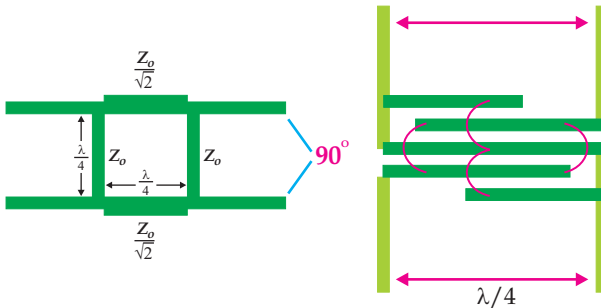
Figure 10.7. (a) Branch-line coupler. (b) Lange coupler

implementation is the integration of directional couplers. Couplers are often implemented as quarter wavelength transmission lines in microstrip, stripeline, or waveguides form. Branch-line hybrid coupler (Figure 10.7(a)), and Lange couplers (Figure 10.7(b)) are two popular hybrid coupler implementations [22]. As shown in Figure 10.7, they both depend on the realization of $\lambda/4$ lines. While

on-chip $\lambda/4$ line is 8.7mm at 5GHz band, which is completely impractical for silicon integration, it drops to 1.6mm at 24GHz which can be realized. In the original design in [22], the matching network and the couplers are designed using a low temperature co-fired ceramic LTCC material and integrated in the package. Other publications uses lumped element representation of directional couplers as capacitor-inductor-capacitor CLC π-networks [24] which has limited bandwidth compared to transmission-line implementation. For mm-wave frequencies, possible on-chip implementation of directional couplers has recently been reported. In [25], a 30GHz branch-line coupler is implemented using thin film microstrip lines (TFMS) in SiGe HBT technology. The 30GHz on-chip coupler has -4dB insertion loss which can largely affect the power amplifier performance. It is worth noting that for the specific case of power amplifiers, the width of the microstrip lines used in the coupler has to be wide enough to handle the large currents in the output stage. The area on chip of the directional coupler is also considerable compared to other passives. At 30GHz, the total area of the directional coupler is 1.25mm$\times$1.25mm.

4.2   Transformer Coupled Power Combining Balun

Aoki et al. [26] have proposed the distributed active transformer DAT technique which functions as an N-way power combiner. Figure 10.8 shows a conceptual view of this technique. The power requirement in this case is distributed among a number of differential driver stages whose outputs are connected in series through the distributed transformer. In the simplified layout shown, each center-tapped drain is the primary of an on-chip transformer that is realized out of coupled lines. The secondary is a one-turn square inductor, where each arm is coupled to its corresponding center-tapped primary. Because the four arms of the secondary are connected in series, the voltage contribution adds as desired and produces the output voltage across the load R_L. The impedance transformation implies that the current in the secondary is less than that of the primary thus narrower lines can be used for the secondary.

Generalizing to N differential pairs with N output transformers, the voltage will be boosted $2N$ times with a power boost of $4N^2$. The value of inductances in the primary and secondary of the transformer is chosen to satisfy maximum quality factor conditions to minimize the losses in the power combiner. Using a 4-way distributed transformer, the value of primary inductance will be equal to $1/16$ of the secondary inductance. This translates into very low inductance value for the primary of the transformer. If spiral inductors/transformers on a silicon substrate are to be used, the small primary inductance results in extremely short metal lines. Inter-winding these short primary metal lines with the multi-turn secondary force them to be very narrow. Unfortunately, this reduces the Q of both primary and secondary circuits significantly. This is the main motivation behind using the coupled lines or slab inductors shown in Figure 10.8.
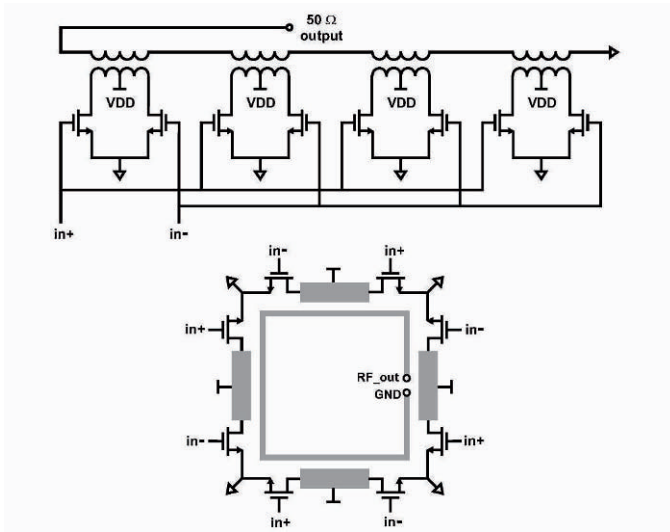
Figure 10.8. Conceptual schematic and layout of a 4-stage PA coupled through a distributed active transformer [26].

Although this technique is by far one of the most elegant ways for output power boosting using low breakdown voltage transistors, the performance is also limited by the losses in the power combining network and practical load values seen by each stage. The imperfect magnetic coupling between the primary and secondary coils increases the loss in the power combiner. Monolithic transformers typically use multi-turn primary and secondary coils to achieve a high magnetic coupling ratio ($K > 0.8$). For a single turn planar transformer as the one used in [26], the K-factor is limited to 0.6 even when minimal metal spacing is used (this is also an issue since top metal layers are usually spaced further than lower metal layers and metal-metal spacing is determined by the technology design rules). Higher coupling factor transformers have been recently proposed in [11] that can synthesize small load impedances and supply sufficient DC current to bias the transistors.

5. Case Study: Design of a 24GHz Power Amplifier Based on Distributed Active Transformer

Using a 4-way power combiner as shown in Figure 10.8 which has a 1:1 transformation ratio between the total primary (4 sections) and secondary in order to simplify the practical realization on chip, the 50Ω load impedance is converted to 12.5Ω seen by each amplification stage. The design can then be divided into;

- the design of a power combiner using slab inductors on chip,

- the design of an output stage with a 12.5Ω load to deliver 100mW output power(total power=400mW).

5.1   Slab Inductor Design and Characterization

Slab inductors are simply metal lines on top of silicon substrate. They can be used for low inductance value implementation and offer much higher Q than spiral inductors, as well as lower area occupation. For the design of the power amplifier combining network, we need an analytical slab inductor optimization flow to determine the value of inductance that provides the highest quality factor at the frequency of interest and satisfy the current handling requirements for the primary of the transformer.
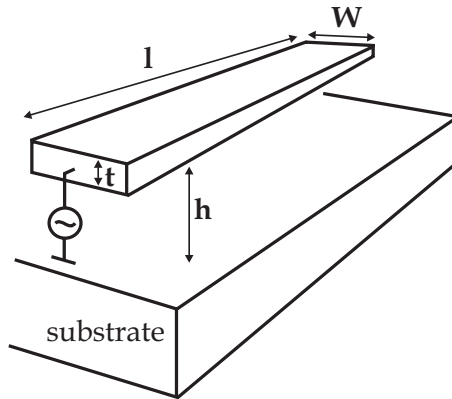
Figure 10.9.   A generic slab inductor structure.

A generic slab inductor structure is shown in Figure 10.9, where the design parameters are

- W = the width of the conductor.

- l = the length of the conductor.

- t = the thickness of the conductor

- h = the distance between the substrate and the conductor

The desired inductance value and the quality factor are a function of these parameters as well as the substrate resistivity.

- Selection of Metal Layer (*t* and *h*): Advanced silicon technologies provide multiple metal layers. In the used technology, we have 7 metal layers with the top being a thick metal layer. Different metal layers have different

thicknesses and are located at different distances from the ground. The quality factor is proportional to t and h. In the process used, the 7th metal layer is the top metal which has the highest distance from the substrate and the highest thickness. In a traditional design, the top metal layer would be the first choice, however, according to the technology physical design rules, the spacing between parallel top-metal lines should be greater than few microns which will severely degrade the coupling coefficient of the transformer down to 0.4. The same design rules are valid for the 6th metal of the process, thus it is also not practical for transformer realization. For this reason, the 5th metal is selected for the implementation of the slab inductors. Although its thickness is 0.5μm, the transformer implemented using this layer has a coupling coefficient close to 0.6.

- Inductor sizing(W and l): In this design, the main concern is the maximization of the quality factor while preserving a width that satisfies current electro-migration rules for the output stage. We used first-order equations for inductance calculation while final design parameters are determined using electromagnetic simulations.

  - Quality factor: We start with the simple quality factor formula that neglects the substrate resistance and the capacitive coupling between the metal line and the substrate. While this is a valid assumption at lower frequencies, it is not valid for final calculations at mm-wave frequencies. This was just used to provide a first cut design that will later be modified according to EM simulations. The following simple formula can be used to find the quality factor.

  $$Q = \frac{L\omega}{R} \tag{10.4}$$

  - Inductor dimensions: As a starting point, some sample inductance values proper for slab inductor realization are selected, such as 50pH, 100pH, 150pH.... Several set of design parameters (W and l for constant conductor thickness) give the same inductance value. The dimensions of the slab inductors are calculated using the following equation [27]:

  $$L_{rect} = \frac{l\mu}{2\pi}\{ln\frac{2l}{W+t} + 0.5 + \frac{\sqrt{W^2 + t^2 + 0.46tW}}{3l} - \frac{W^2 + t^2}{24l^2}\} \tag{10.5}$$

  where μ is the magnetic permeability of free space ($\mu = 4\pi\ 10^{-7}$). The different dimensions that generate the same inductance value have different parasitic resistances. The series parasitic resistance is calculated using the equation:

Table 10.1.   Example of inductance values and corresponding dimensions.

The sample inductance values	The selected sample widths
50pH(75pH)	30μm,40μm,50μm,60μm
100pH(130pH)	30μm,40μm,50μm,60μm
200pH(235pH)	30μm,40μm,50μm,60μm

$$R = \frac{L}{W} R_\square \qquad\qquad (10.6)$$

where $R_\square$ is the sheet resistance of the conductor. Once the dimensions are determined using equation 10.5 and the parasitic resistance is calculated according to equation 10.6, the initial approximate value for the quality factor of each inductor is determined using equation 10.4.

- Table 10.1 shows the sample inductance values and selected conductor widths. For each inductance value, the width is increased to reduce the parasitic resistance. The length of the inductor is changed each time to obtain the same inductance value. The initial values are obtained using equations 10.4, 10.5 and 10.6. The lower boundary of the width is determined by the current handling capability of the metal line or the parasitic resistance of the metal line. For the upper boundary, if the width is selected larger than 60μm, either the aspect ratio is close to 1 or the length of the conductor is too high depending on the target inductance value.

- Note that equation 10.4 completely ignores all the loss sources in the inductor except the series resistance of the metal. The electromagnetic simulation results take into account the substrate and skin effects which can be significant at mm-wave frequencies. Also equation 10.5 gives very accurate results when the dimensions of the inductor are sufficiently large. However, for the inductors designed, the dimensions are so small that the results of equation 10.5 diverge from EM simulation results. The inductors were simulated in an EM simulation environment (SONNET) for different set of design parameters. The quality factor of each structure at 24GHz is simulated as shown in Figure 10.10. Note that while equation 10.5 gives the value of the inductance in terms of (W,l,t), the effective slab inductance determined using EM simulations is higher. The actual inductance values are reported between brackets in Table 10.1.

In Figure 10.10, the feasible inductance values for the highest Q can be determined. For the 100pH(130pH), the value of Q saturates as the width increases. The Q curve for the 200pH(235pH) is much lower than the
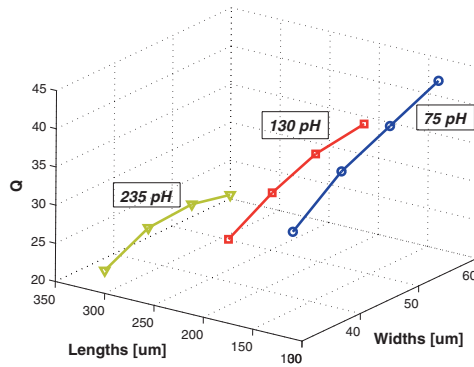
Figure 10.10. EM Simulations showing the quality factor for different inductance values.

other two cases. The lower boundary for the inductor is determined by practical limits. Inductance values lower than 50pH(75pH) may not be realized precisely since it is comparable to interconnect inductances. They may also be sensitive to process variation and/or suffer from the absolute tolerance of the used design tools.

Final design dimensions of the inductor are determined by considering the trade-offs that are stated above. An inductance of 80pH is selected as the final optimum value. The width is selected as $50\mu m$ and the length of the inductor is $175\mu m$. For this width, the metal can handle an rms current slightly higher than 200 mA, which is equal to two times of the rms operational current for this example. The inductor has a Q of 37 at 24GHz.

- The simulated S-parameters of the inductor are used to create an equivalent R-L-C-K netlist for the Spectre simulations, where K is the coupling coefficient. Figure 10.11 shows the 6-segment equivalent circuit used for the transformer and Table 10.2 shows the corresponding design parameters.

5.2   Single-stage Amplifier Design

The basic differential push-pull power amplifier stage is shown in Figure 10.12. Each unit stage provides a quarter of the required output power. Using the inductance value determined from the slab inductor analysis, the devices size (m1 and m2) are changed such that the required output power is obtained at the secondary winding. The swing at the output of the primary side is determined by the real impedance at this point and the transient current that is supplied by the driving transistors. Cl is used to cancel the inductive part of the impedance of primary winding at the frequency of operation. The
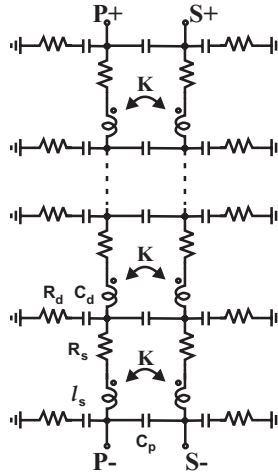
Figure 10.11.   E6-segement equivalent circuit for modeling the distributed transformer.

Table 10.2.   Equivalent circuit component values.

Component	value
R_d	100Ω
C_d	3fF
R_s	70Ω
L_s	18pH
C_P	$10fF$
K	0.6
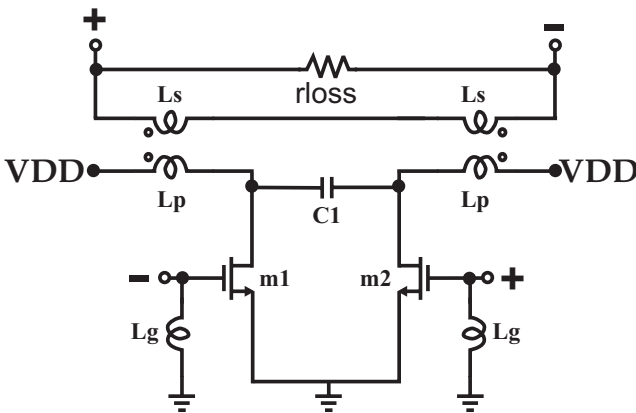
Figure 10.12.   Basic Differential Stage.

Table 10.3. Final Design Parameters.

Component	value
VDD	2V
Lp	80pH
Ls	80pH
Lg	30pH
$(W/L)_{m1,m2}$	200x (2μm/0.18μm)

real output impedance is determined by the output resistance of transistors, loss resistance of primary winding and transferred loss resistance of secondary winding. Estimating the size of the transistors theoretically is not feasible due to the lack of accurate large signal models for the devices. Instead, we changed the size of the active device in simulations to obtain the required output power at the secondary winding. However, if the device size is changed, the parasitic capacitance of the wide driver transistor will also change and the value of Cl should be readjusted to provide resonance condition at the operating frequency. Lg is used to tune out the gate capacitance. Final design parameters are given in Table 10.3.

5.3   Simulation Results

Figure 10.13 shows the power combination process at a given output power level. The signal adds up in the secondary windings of the transformer. The power amplifier is simulated using 0.18μ CMOS technology parameters. It can deliver a 1-dB compression point of 23.3dBm, a small signal gain of 18dB with output saturation power of 25.5dBm and maximum drain efficiency of 35.6% using a 1.8V supply as shown in Figure 10.14.

The power amplifier using magnetically coupled transformers has a broadband characteristic due to the transformer action as shown in Figure 10.15. It is worth noting that the same technique has been recently used for the implementation of a 21-26GHz three stage common-base power amplifier in SiGe technology[11]. The design of the transformer in the power combiner was optimized for higher coupling coefficient (k=0.9) and double the turns ratio by using a self-shielded structure as shown in Figure 10.16 which resembles a coaxial transformer. The self shielding design technique was applied to all on-chip passive devices to optimize inter-stage and I/O coupling at mm-wave frequencies while supplying the DC current ratings for the PA. 23dBm output power with 20% power added efficiency at 22GHz and 13% at 24GHz were measured using a 1.8V supply.
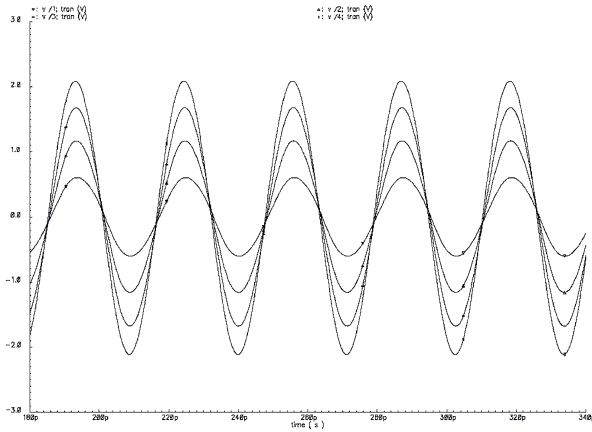
Figure 10.13.   Power combining in time domain.
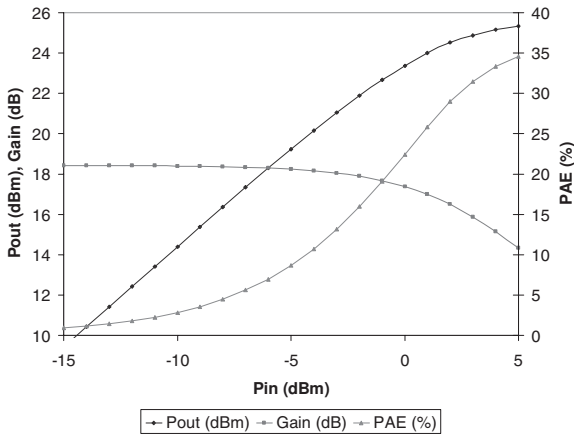
Figure 10.14.   Pout, Gain and PAE v.s. Pin.

6.   Summary and Conclusions

Wireless communications beyond 20GHz are gaining momentum as they offer several advantages compared to 1-6GHz regime. The reduction in an-
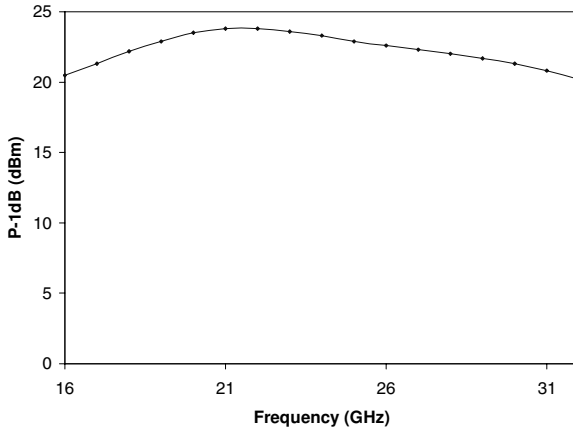
Figure 10.15. Simulated 1-dB compression point versus frequency.
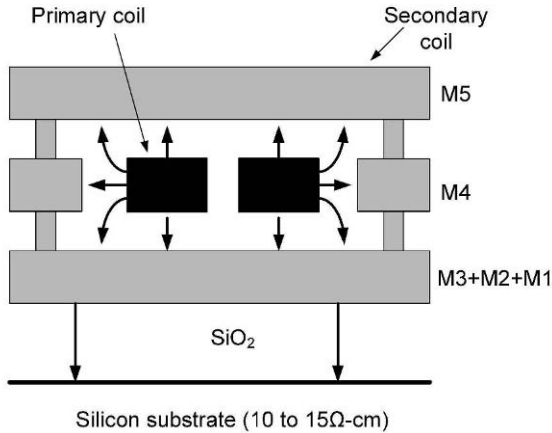
Figure 10.16. Self-shielded monolithic transformer used in the distributed-active-transformer DAT-based power amplifier reported in [11].

tenna size at mm-wave frequencies makes compact, multiple antenna arrays more practical. Electronic beam steering capabilities of these short wavelength arrays antennas will simplify their deployment in WLAN applications. These applications will be available in the future at much lower costs thanks to the advances in CMOS and SiGe technologies. Although these advances will gen-

generate faster devices, the breakdown voltage of these devices will dramatically decrease, limiting their capabilities for power amplifier implementation and thus complete integration with the communication transceiver. We have presented different techniques for power amplifier design in the mm-wave range that explores circuit ideas to overcome the low breakdown-voltage barrier. Power combining techniques through the use of on-chip lumped or distributed transformers, directional couplers are possible ways to generate higher power from multiple amplifier stages. Realizing high quality passive elements to be used in these combiners is the primary challenge in the mm-wave range. Accurate modeling of these devices is also crucial to obtain first pass designs. Finally, recent publications of power amplifiers in the 24GHz, 60, and 77GHz using CMOS and SiGe technologies show that silicon-based technologies has the potential to compete with III-V technologies for medium power applications in the mm-wave range.

References

[1] Peter Russer, "Si and SiGe Millimeter-Wave Integrated Circuits," *IEEE Trans. On Microwave Theory and Techniques*, vol. 46, no. 5, pp. 590-603, May 1998.

[2] X. Guan, H. Hashemi, and Ali Hajimiri, "A Fully Integrated 24-GHz Eight-Element Phased-Array Receiver in Silicon," *IEEE J. Solid State Circuits*, vol. 39, no. 12, pp. 2311-2320, Dec. 2004.

[3] B. Floyd, et al., "SiGe Bipolar Transceiver Circuits Operating at 60GHz," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp.156-167, Jan. 2005.

[4] A. Babakhani, et al., "A 77GHz 4-Element Phased Array Receiver with On-Chip Dipole Antennas in Silicon," *ISSCC Dig. Tech. Papers*, Feb. 2006, pp. 180-181.

[5] C. H. Doan et al., "Millimeter-Wave CMOS Design," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 144-155, Jan. 2005.

[6] A. Natarjan, et al., "A 77GHz Phased Array Transmitter with LO-Path Phase-Shifting in Silicon," *ISSCC Dig. Tech. Papers*, Feb. 2006, pp. 182-183.

[7] C. Wang, et al., "A 60GHz Transmitter with Integrated Antenna in 0.18μ SiGe BiCMOS Technology,"*ISSCC Dig. Tech. Papers*, Feb. 2006, pp. 186-187.

[8] L. M. Franca-Neto, B. A. Bloechel, and K. Soumyanath, "17GHz and 24GHz LNA designs based on extended S parameters with micro-strip-on-die in 0.18um logic CMOS technology," in *Proc. Eur. Solid-State Circuits Conf.*, Sept. 2003, pp. 149-152.

[9] I. Gresham and J. Jekins,"A low-noise broadband SiGe mixer for 24GHz ultra-wideband automotive applications," *IEEE Radio and Wireless Conf.*, Aug. 2003, pp. 361-364.

[10] S. Y. Yue, D. Ma, and J.R. Long,"A 17.1-17.3-GHz Image Reject Downconverter with Phase-Tunable LO using 3x Subharmonic Injection Locking," *IEEE J. Solid-State Circuits,* vol. 39, no. 12, pp. 2321-2332, Dec. 2004.

[11] T. S. D. Cheung and J.R. Long,"A 21-26GHz SiGe Bipolar Power Amplifier MMIC," *IEEE J. Solid-State Circuits,* vol. 40, no. 12, pp. 2583-2597, Dec. 2005.

[12] C. Yoo, Q. Huang, "A common-gate switched, 0.9W class-E power amplifier with 41% PAE in 0.25μm CMOS," *VLSI circuits Symposium Digest*, June 2000, pp 56-57.

[13] T. Kuo, B. Lusignan, "A 1.5-W class-F RF power amplifier in 0.25-μm CMOS technology," *ISSCC Dig. Tech. Papers*, 2001, pp 154-155.

[14] T. Sowlati, D. Leenaerts, "A 2.4-GHz 0.18-μm CMOS Self-Biased Cascode Power Amplifier," *IEEE Journal of Solid-State Circuits*, vol. 38, no. 8, August 2003, pp 1318-1324.

[15] K. Choi, D. Allstot, V. Krishnamurthy, "A 900MHz GSM PA in 250nm CMOS with Breakdown Voltage Protection and Programmable Conduction Angle," *IEEE Radio Frequency Integrated Circuit Symposium*, 2004, pp 369-372.

[16] J. Weng and M. Hella, "A Highly-Linear 5-GHz SiGe Power Amplifier with Breakdown-Voltage Protection", Submitted to *Bipolar/BiCMOS Circuits and Technology Meeting*

[17] A. Kmoijani and A. Hajimiri, "A 24GHz, +14.5dBm Fully Integrated Power Amplifier in 0.18μm CMOS," *IEEE Custom Integrated Circuits Conference*, 2004, pp. 561-564.

[18] A. Komijani, et al.,"A WideBand 77GHz, 17.5dBm Power Amplifier in Silicon," Proc. *IEEE CICC*, pp. 571-575, Sept. 2005.

[19] W. Bakalski, et al.,"A Fullt Integrated 7-18GHz Power Amplifier with On-Chip Output Balun in 75GHz- f_T SiGe Bipolar," *IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, Sept. 2003, pp. 61-64.

[20] Peter Vizmuller, *RF Design Guide - Systems. Circuits, and Equations*, Artech House, first edition, 1995.

[21] W. Bakalski, W. Sinihurger, H. Knapp, and A.L. Scholtz, "Lumped and Distributed Lattice-type LC Baluns," *in Proceedings of IEEE International Microwave Symp.*, Seattle, June 2002.

[22] D. Pozar, *Microwave Engineering*, 2nd ed., John Wiley, 1998.

[23] N. Tanzi, "A 1-Watt Doubly Balanced 5GHz Flip-Chip SiGe Power Amplifier," *IEEE Radio Frequency Integrated Circuit Symposium*, 2003, pp 141-144.

[24] D. Ozis, and D. Allstot, "A CMOS 5GHz Phase-Compensated Quadrature Coupler, " *IEEE Radio and Wireless symposium (RWS)* , Jan. 2006, pp. 51-54.

[25] J. Lee, Y. Tretiakov, J. Cressler and A. Joseph, "Design of a Monolithic 30GHz Branch Line Coupler in Sige HBT technology Using 3-D EM Simulation," *2004 Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems*, pp. 274-277.

[26] I. Aoki, et al., 'Fully Integrated CMOS Power Amplifier Design Using the Distributed Active-Transformer Architecture', *IEEE Journal of Solid-State Circuits*, March 2002, pp 371-383.

[27] S. Mohan, "The Design, Modeling and Optimization of On-Chip Inductor and Transformer Circuits," Ph.D Thesis, December 1999, Stanford University.

Chapter 11

MONOLITHIC INDUCTOR MODELING AND OPTIMIZATION

Niklas Troedsson and Henrik Sjöland

It has become a necessity to use on-chip inductors in radio frequency integrated circuits. Particularly oscillators need inductors to archive high performance. In LC oscillators the quality factor of the inductor is critical to the phase noise performance, and since the self-resonance frequency of the inductor will limit the operating frequency and tuning range of the oscillator, careful optimization is needed. As the transistors are scaled to smaller geometries, the supply voltage must be reduced. Since the inductors have almost zero DC voltage drop they can be used to increase the voltage headroom in low voltage RF circuits. For cross-coupled differential pair LC oscillators, an inductor at the source node of the differential pair will not only increase the signal headroom [1], but if resonating at twice the oscillation frequency it will also increase the phase noise performance significantly [2]. The gain and noise factor of LNAs and mixers can also be improved by using inductors to tune out parasitic capacitances [3].

Since both the industry and the market push for higher operating frequencies, it has become increasingly important to estimate self-resonance frequencies accurately. An inductor is needed that not only meets the requirements of the inductance, L, and the quality factor, Q, but also has a minimum amount of capacitance, i.e., has the highest self-resonance frequency possible. An accurate knowledge of the self-resonance frequency is necessary in order to create an optimized design that takes capacitance into account. A time-consuming optimization process is typically needed to find an inductor geometry that meets all these requirements.
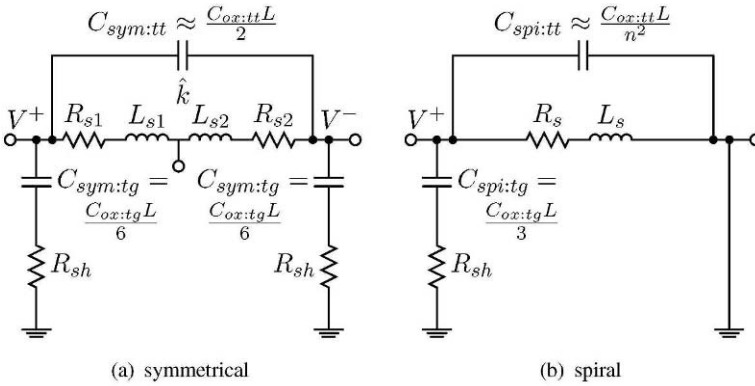
Figure 11.1. Inductor π-model. $C_{ox:tg}L$ is the total capacitance from the inductor trace to ground and $C_{ox:tt}L$ is the total turn-to-turn capacitance.

There is a demand of electromagnetic (EM) simulators that can rapidly extract L, Q (LQ) and self-resonance frequency; and also an accurate inductor model to be used in circuit simulators such as Spice and SpectreRF. This motivated the development of the tool called Indentro [4], which can be used to quickly find an optimized inductor geometry, and to extract a lumped π-model for use in circuit simulators.

1.   Monolithic Inductor Modeling

By modeling monolithic inductors with empirical [5, 6] or semi-empirical formulas [7], computation time is greatly reduced compared to field solvers such as ASITIC [8], FastHenry [9], ADS Momentum, etc. The lumped inductor π-model, Fig. 11.1, is commonly used in circuit simulators. The inductance and series resistance including skin effect can be calculated using concise formulas. Other aspects such as eddy and proximity effects, must be evaluated by electromagnetic field simulators for accuracy [8, 9]. Complex formulas and inductor π-models have, however, been presented for calculating both eddy-current losses over conductive substrates [10] and current crowding (proximity effects) [11], but unfortunately there are not yet any simple formulas to fully describe the properties of integrated inductors, and field solvers are too slow for optimization.

Figure 11.2 and Fig. 11.3 show typical inductor layouts using top metal layer and a crossing (bridge) metal layer. The top metal is often thicker than other metal layers reducing the series resistance. Symmetrical inductors have higher quality factors than spiral inductors, and also require less area than using two spiral inductors in a differential circuit, whereas spiral inductors have higher self-resonance frequencies. Moreover, the more circular an inductor is, the

(a) square   (b) hexagon   (c) octagon   (d) circle
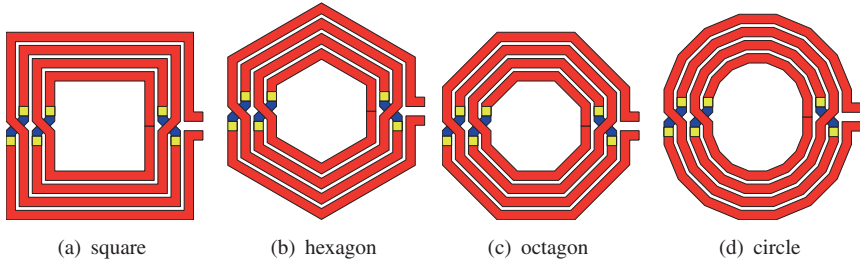
Figure 11.2.   Layout of symmetrical inductors.

(a) square   (b) hexagon   (c) octagon   (d) circle
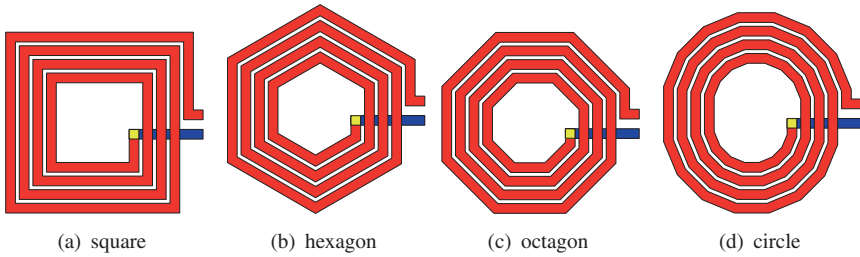
Figure 11.3.   Layout of spiral inductors.

higher the self-resonance frequency and quality factor. A circular geometry has the shortest length for a given enclosed area, hence the least capacitance and series resistance for a fixed inductance. The performance therefore increases going from left to right in Fig. 11.2 and Fig. 11.3. However, the improvements beyond octagonal shape are very small.

In [12] a fully symmetrical inductor is presented that combines the geometry of spiral and symmetrical. The proposed inductor has a lower Q and inductance, but instead has the advantage of better immunity to environmental noise. A new method to enhance the quality factor of inductors by suppressing current crowding has been presented in [13]. The idea is to have 2 or 4 narrow paths in parallel (as opposed to 1 wide for conventional inductors), and let these paths change place 2 or 4 times between the inner and the outer turns. A 40% improvement in Q was reported using a 0.18μm CMOS process.

We will now present how to easily calculate the desired inductance value as well as the different unwanted parasitics, that is, capacitances and resistances. To achieve the best possible accuracy estimating the self-resonance frequency, the capacitances are calculated using co-planar microstrip theory and the principle of conservation of energy.
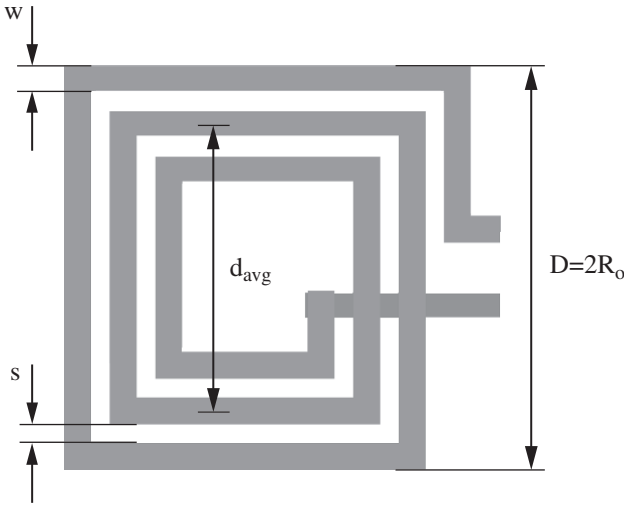
Figure 11.4. Layout of spiral inductors, showing design parameters.

1.1   Inductance

Simple expressions for the inductance of monolithic inductors are presented in [14]. They include three models: Modified Wheeler, Current Sheet Approximation, and Data Fitted Monomial Expression. All three have been compared to over 19,000 inductors simulated with ASITIC [8]. The models use a filling factor defined by (11.1). The average diameter of the inductors is (11.2), Fig. 11.4:

$$d_{rho} = \frac{2\,R_o - 2\,(R_o - w - (n-1)(w+s))}{2\,R_o + 2\,(R_o - w - (n-1)(w+s))} \tag{11.1}$$

$$d_{avg} = \frac{2\,R_o + 2\,(R_o - w - (n-1)(w+s))}{2} \tag{11.2}$$

where n is the number of turns. The parameters R_o, w, and s are in μm, which yields d_{avg} in μm; d_{rho} becomes dimensionless.

Modified Wheeler

In 1928, during the early days of radio, Wheeler [15] presented several formulas for discrete planar spiral inductors. The formulas in [14] are modified to apply to monolithic planar spiral inductors (11.3):

$$L_{mw} = K_1\,\mu_0\,n^2\,\frac{d_{avg}}{1 + K_2\,d_{rho}} \cdot \frac{1}{10^{-9}} \tag{11.3}$$

where K_1 and K_2 are geometry-dependent constants, Table 11.1. L_{mw} will be in nH when d_{avg} is in m.

Table 11.1. Constants: Modified Wheeler.

Geometry	K_1	K_2
square	2.34	2.75
hexagon	2.33	3.82
octagon	2.25	3.55

Table 11.2. Constants: Geometric mean distance (GMD).

Geometry	c_1	c_2	c_3	c_4
square	1.27	2.07	0.18	0.13
hexagon	1.09	2.23	0	0.17
octagon	1.07	2.29	0	0.19
circular	1.00	2.46	0	0.20

Current Sheet Approximation

The current is approximated to flow at the sides of the conductor in sheets with equal current density. After evaluation using the concept of geometric mean distance (GMD) and arithmetic mean distance (AMD), the resulting expression becomes (11.4). The maximum error is 8% for $s \leq 3w$. Most inductors are built with $s \leq w$, since the magnetic coupling is stronger with smaller spacing. A large spacing is only desirable to reduce interwinding capacitance.

$$L_{gmd} = \mu_0 \, n^2 \, d_{avg} \, \frac{c_1}{2} \left(\log \left(\frac{c_2}{d_{rho}} \right) + c_3 \, d_{rho} + c_4 d_{rho}^2 \right) \frac{1}{10^{-9}} \qquad (11.4)$$

where c_1, c_2, c_3, and c_4 are geometric constants, Table 11.2. L_{gmd} will be in nH when the factor d_{avg} is in m.

Data Fitted Monomial Expression

The final expression (11.5) is based on a data-fitting technique. The coefficients of this expression are developed from a library of inductors [14], but also seem to agree well with results from different simulators:

$$L_{mon} = \beta \, (2 \, R_o)^{\alpha_1} \, (w)^{\alpha_2} \, (d_{avg})^{\alpha_3} \, (n)^{\alpha_4} \, (s)^{\alpha_5} \qquad (11.5)$$

where the coefficients are geometrically dependent, Table 11.3. L_{mon} will be in nH when constants R_o, w, s, and d_{avg} are in μm.

Table 11.3. Constants: Data-fitted monomial expression.

Geometry	β	α_1	α_2	α_3	α_4	α_5
square	$1.62 \cdot 10^{-3}$	-1.21	-0.147	2.40	1.78	-0.030
hexagon	$1.28 \cdot 10^{-3}$	-1.24	-0.174	2.47	1.77	-0.049
octagon	$1.33 \cdot 10^{-3}$	-1.21	-0.163	2.43	1.75	-0.049

1.2 Resistance: Metal Losses

The ohmic losses in the windings are caused by the finite conductivity of the metal. The largest contribution is typically from the DC resistance (11.6):

$$R_{DC} = \rho \frac{l}{wh} \qquad (11.6)$$

where ρ is the resistivity of the material (Ωm), l is the total length of the windings (m), and w and h are width and height (m) of the winding trace.

As the frequency increases, the skin-effect becomes prominent. A magnetic field is induced in the conductor, forcing the current to flow near its edges. The series resistance is then increased since the current flows through a smaller effective area of the conductor. A formula for high frequency resistance in rectangular conductors is given by (11.7) [16]:

$$R_{AC} = R_{DC} \left[1 + \left(\frac{f}{f_l} \right)^2 + \left(\frac{f}{f_u} \right)^5 \right]^{\frac{1}{10}} \qquad (11.7a)$$

$$f_l = \frac{\pi \rho}{2\mu wh} \qquad (11.7b)$$

$$f_u = \frac{\pi^2 \rho}{\mu h^2} \left[K \left(\sqrt{1 - \frac{h^2}{w^2}} \right) \right]^{-2} \qquad (11.7c)$$

where μ is the permeability [Vs/Am] of the metal. The frequencies f_l and f_u are the boundary frequencies of the low and the high frequency cases, respectively. K is the elliptic integral of the first order (11.8):

$$K(x) = \int_0^{\frac{\pi}{2}} \frac{1}{\sqrt{1 - x^2 \sin^2(\phi)}} \, d\phi \qquad (11.8)$$

where $0 \leq x \leq 1$. The elliptic integral function has been implemented in Matlab, making it easy to use.

The proximity effect can only occur if two or more conductors are present. The magnetic field from each conductor then affects the current flow in the other

ones, resulting in a non-uniform current distribution. The proximity effect is similar, in this respect, to the skin effect. The effective cross-section of the conductor decreases, increasing the resistance. Metal losses in the inductor, such as DC, skin effect, proximity effect, and eddy currents, can be simulated with FastHenry [9].

1.3 Electric and Magnetic Substrate Losses

The eddy current effects due to the nearness of the substrate can be severe, depending on the substrate resistivity and frequency of operation [17]. The higher the resistivity, the higher the Q value and self-resonance frequency. The higher the frequency, the greater the losses. The resistivity can be divided into three regimes [17]: the high region from 10 to $1000\Omega cm$ is the inductor mode regime where metal losses dominate, since the substrate behaves as a dielectric represented by a small oxide capacitance and a resistance large enough to suppress any resonance; the middle region from 0.1 to $10\Omega cm$ is the resonator mode regime where the substrate resistance and capacitance start to resonate with the inductance and drastically lower the Q and f_{sr}; the low region from 0.001 to $0.1\Omega cm$ is the eddy current regime where eddy currents dominate and further degrade Q. Eddy currents in the substrate ruin also the inductance value and cannot be shielded, although direct (capacitively coupled) ohmic substrate losses can.

To avoid eddy currents in the shield, a patterned ground shield with narrow traces is usually used under the inductor [6, 18, 19]. Another approach would be to use as high a shield resistivity as possible [17]. For the rest of this section, only electric substrate losses will be considered, i.e., situations with no shielding, so that shield resistance, R_{sh}, is not present and therefore does not need to be taken into account.

An expression for the substrate resistance is presented by [20]. The current flow is possible due to the capacitive coupling between the traces of the inductor and the substrate (with no shield present) through the non-conductive dielectric (often SiO_2):

$$R_{sub} = \rho \frac{h_{Sub}}{A} \tag{11.9}$$

where A is the effective area underneath the traces, from approximately the inner turn to the outer turn, and h_{Sub} is the substrate thickness.

The capacitance can be calculated using a the plate capacitance approximation with the same parameters as for the resistance:

$$C_{sub} = \epsilon_r \epsilon_0 \frac{A}{h_{Sub}} \tag{11.10}$$

(a) without shield   (b) with shield
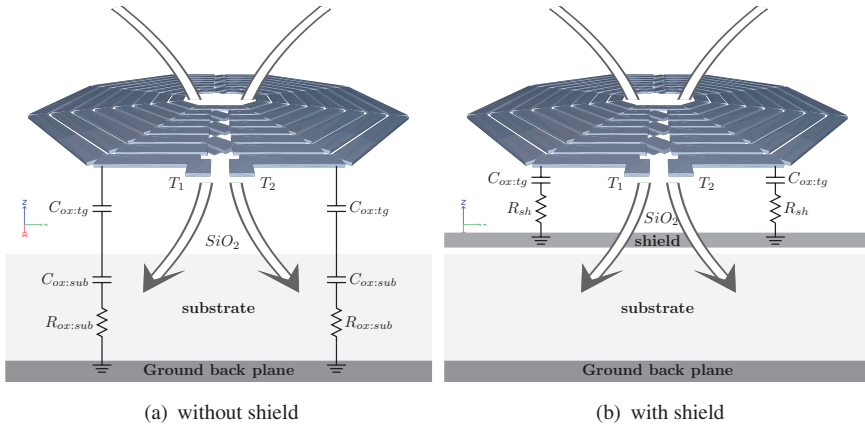
Figure 11.5.   (a) Electric and magnetic substrate losses without shield. (b) Magnetic (eddy) substrate losses with shield.

which results in the expression:

$$R_{sub}C_{sub} = \rho\epsilon_r\epsilon_0 \qquad (11.11)$$

The result can be regarded as the time constant of the substrate material, independent of geometry parameters.

1.4   Capacitance

The basic ideas used to calculate the capacitances of the lumped inductor π-model is:

1 To employ two dimensional co-planar microstrip theory to calculate the distributed capacitance from trace to ground and between inductor traces.

2 To use the principle of conservation of energy to calculate the equivalent lumped capacitances.

The capacitances seen at a terminal of an inductor is dependent on the geometry of the inductor and on the signals applied to both terminals; ground is also an signal. If the terminal voltages are different we assume a linear voltage profile along the inductor traces. The total capacitive energy stored in the distributed capacitances can then be found by integration. The equivalent lumped capacitances at the terminals, storing the same amount of energy, can then be found.

The capacitance theory is presented in [7] and will therefore not be dealt with at any length here. Fig. 11.6 depicts capacitances that can be calculated by
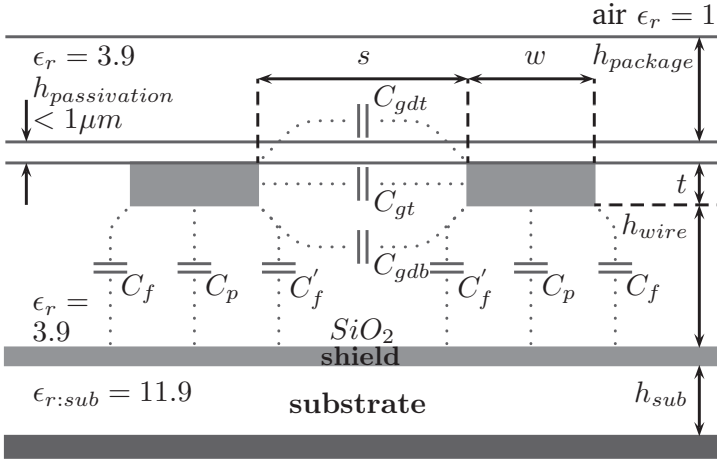
Figure 11.6. Cross-section of microstrips on-chip and their capacitances.

common microstrip theory, slightly attuned to fit our monolithic environment. Using the microstrip theory we extend the capacitance analysis to include fringe capacitances, C_f, in addition to the ordinary plate capacitance, C_p, which will significantly improve the accuracy of the capacitance estimation.

The distributed capacitances of inductors were analyzed by investigating the capacitances of co-planar microstrips [21, 22]. From basic microstrip theory we derived equations fitting the surroundings of on-chip wires, i.e. the wires are completely surrounded by an isolation material (SiO_2) and rest on top of a substrate (Si) as shown in Fig. 11.6. This is an approximation of the on-chip isolation material and the package material usually directly on top of the chip, all here assumed to have the same relative permittivity ϵ_r. The Equations (11.12)-(11.15) slightly differs from [21] due to these assumptions.

Fig. 11.6 depicts the capacitances from odd-mode excitation of parallel coupled microstrip lines. Odd-mode corresponds to the case when one strip is charged positively and the other negatively. This is the case of maximum energy storage between the strips. The capacitances (per length) are summarized in (11.12), where C_p is the plate capacitance, C_f the fringe capacitance in absence of an adjacent strip, and C'_f is the fringe capacitance in presence of such a strip. C_{gdb} is the fringe capacitance between the edges at the bottom of the strips, and C_{gdt} the top.

$$C_p = \epsilon_0 \epsilon_r \frac{w}{h} \tag{11.12a}$$

$$C_f = \frac{1}{2} \left(\frac{\epsilon_r}{c\, Z_{01}} - C_p \right) \tag{11.12b}$$

$$C_f' = \frac{C_f}{1 + A \dfrac{h}{s} \tanh\left(8 \dfrac{s}{h}\right)} \tag{11.12c}$$

$$C_{gdt} = \frac{1}{2} \epsilon_0 \epsilon_r \frac{K(k_{fa}')}{K(k_{fa})} \tag{11.12d}$$

$$C_{gdb} = \frac{1}{2} \frac{\epsilon_0 \epsilon_r}{\pi} \ln\left(\coth\left(\frac{\pi s}{4h}\right)\right) + 0.013 \frac{C_f}{s/h} \tag{11.12e}$$

where Z_{01} is the characteristic impedance of the microstrip with ϵ_r of the isolation material set to 1 (single microstrip in air), (11.13). The effective width w_e is given by (11.14).

$$Z_{01} = \begin{cases} 60 \ln\left(8\dfrac{h}{w_e} + \dfrac{w_e}{4h}\right) & \text{if } \dfrac{w}{h} \le 1, \\[2ex] 120\pi \left(\dfrac{w_e}{h} + 1.393 + \right. \\ \left. 0.667 \left(\dfrac{w_e}{h} + 1.444\right)\right)^{-1} & \text{if } \dfrac{w}{h} \ge 1. \end{cases} \tag{11.13}$$

$$w_e = \begin{cases} w + 5\dfrac{t}{4\pi}\left(1 + \ln\left(4\pi\dfrac{w}{t}\right)\right) & \text{if } \dfrac{w}{h} \le \dfrac{1}{2\pi}, \\[2ex] w + 5\dfrac{t}{4\pi}\left(1 + \ln\left(2\dfrac{h}{t}\right)\right) & \text{if } \dfrac{w}{h} \ge \dfrac{1}{2\pi}. \end{cases} \tag{11.14}$$

Furthermore, the variables A, k_{fa}, and k_{fa}' are given by (11.15a), (11.15b), and (11.15c) respectively.

$$A = \exp\left(-0.1 \exp\left(2.33 - 2.53 w/h\right)\right) \tag{11.15a}$$

$$k_{fa} = \frac{s/h}{s/h + 2w/h} \tag{11.15b}$$

$$k_{fa}' = 1 - k_{fa}^2 \tag{11.15c}$$

$K(m)$ is the elliptic function of the first kind,

$$K(m) = \int_0^{\frac{\pi}{2}} (1 - m \sin^2 \theta)^{-\frac{1}{2}} d\theta \tag{11.16}$$

where m is limited to $0 \leq m \leq 1$. The elliptic functions are implemented in Matlab for easy use.

The sidewall plate capacitance C_{gt} between wires is given by (11.17).

$$C_{gt} = \epsilon_0 \epsilon_r \frac{t}{s} \qquad (11.17)$$

To facilitate a comparison of the fringe capacitances and the plate capacitances we define capacitance ratios for the following cases:

- (11.18a): a single wire, trace-to-ground.

- (11.18b): the outer conductor in an array of co-planar parallel conductors, trace-to-ground.

- (11.18c): the mid-conductor in an array of three conductors, trace-to-ground.

- (11.18d): capacitance between two conductors relative sidewall plate capacitance, trace-to-trace.

$$C_{1:tg} = \frac{C_p + 2C_f}{C_p} \qquad (11.18a)$$

$$C_{2:tg} = \frac{C_p + C_f + C'_f}{C_p} \qquad (11.18b)$$

$$C_{3:tg} = \frac{C_p + 2C'_f}{C_p} \qquad (11.18c)$$

$$C'_{gd} = \frac{C_{gt} + C_{gdb} + C_{gdt}}{C_{gt}} \qquad (11.18d)$$

In Fig. 11.7 the capacitance ratios (11.18a)-(11.18c) are plotted. The process parameters used are $t = 3\mu m$ and $h = 6\mu m$ (thick top metal). The results are interesting since only a few earlier papers, e.g. [23, 24], have included the fringe capacitance in their models, but solely for optimization of spiral inductors. As will be described in 1.5 the symmetrical inductors are much more sensitive to the turn-to-turn capacitances, so good modeling of all the capacitances is necessary, in particular of the fringe capacitances.

As can be seen, for most commonly used wire geometries and especially for small wire widths, the fringe capacitances dominate over the plate capacitances. Obviously, the fringe capacitances must be taken into account for an accurate estimation of the parasitics and the self resonance frequency.

The ratio $C_{1:tg}$ is spacing independent and represents the maximum trace to ground ratio value that can be obtained (infinite spacing). For wide wires, about $30\mu m$, this ratio is around 2.25, meaning that the total capacitance is 2.25 times
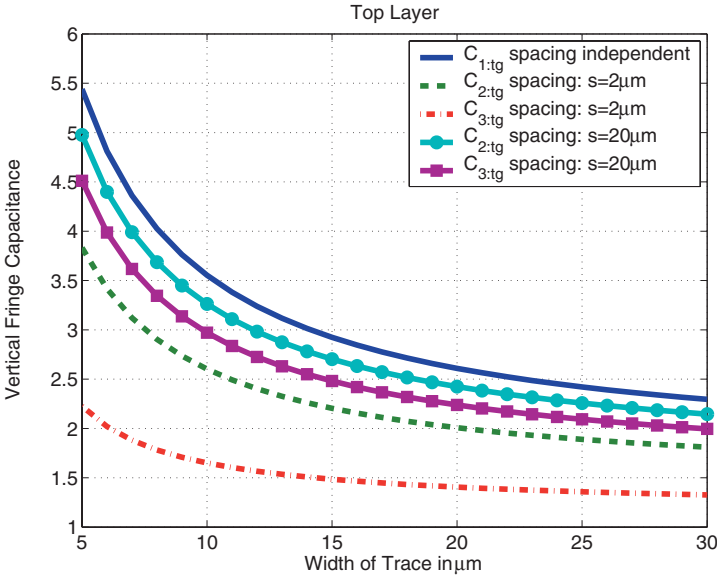
Figure 11.7.   Capacitance ratios versus width, w. ($t = 3\mu$m, $h = 6\mu$m.)

larger than the plate capacitance. As wires become narrower, at $5\mu m$ the ratio rises to 5.5.

The ratios $C_{2:tg}$ and $C_{3:tg}$ are spacing dependent. As can be seen, for large spacing between wires they approach the $C_{1:tg}$ ratio, i.e. the roof. For smaller spacings the ratios $C_{2:tg}$ and $C_{3:tg}$ differentiate more clearly. $C_{3:tg}$ is always smaller than $C_{2:tg}$.

The sidewall ratio C'_{gd} (11.18d) is presented in Fig. 11.8. The simulations have been done for $10\mu m$ wide conductors. First observe the thick top metal that has 2.5-6 times the sidewall plate capacitance when the spacing is swept from $2\mu m$ to $20\mu m$. The thin top metal has an even larger ratio of 5 to 20 times. This clearly shows that the turn-to-turn capacitance is completely dominated by fringing, and calculating it using plate capacitance only results in gross under estimations.

1.5   The Equivalent Lumped Capacitances

The theory of finding equivalent lumped representations of the distributed capacitances will only be briefly addressed here, since it is also presented in [7]. To find the equivalent capacitances of different inductor topologies, the principle of conservation of energy is employed. One of the key results of the
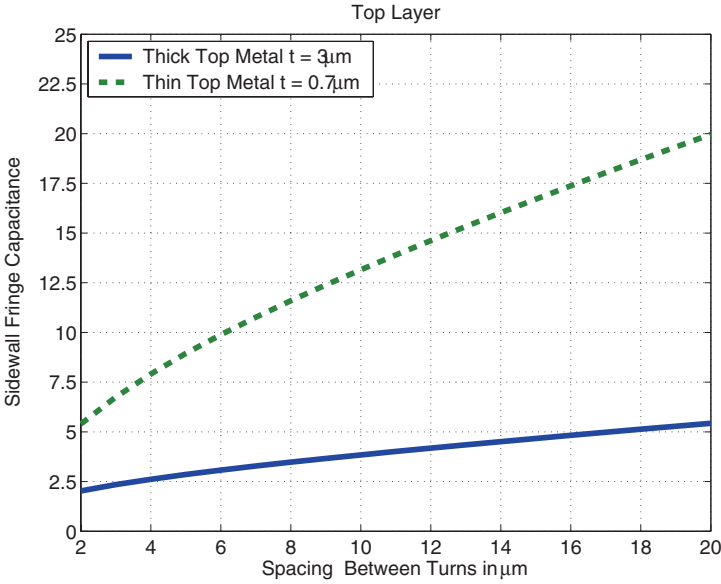
Figure 11.8. Capacitance ratio of C'_{gt} versus spacing between wires, s. ($h = 6\mu$m, and $w = 10\mu$m.)

analysis in [7] is that the turn-to-turn capacitance exerts a major influence on the self-resonance frequency of symmetrical inductors, but has little effect in spiral inductors and can thus often be neglected.

Energy Storage from Trace to Substrate
One Terminal Grounded Spiral Inductor

Through (11.19) the equivalent capacitance seen at the signal node is derived. $W_{ox:tg}$ is the energy stored in the distributed capacitance from trace to ground, which must be equal to $W_{spi:tg}$, the energy of the equivalent capacitance. We have assumed a linear voltage drop along the inductor traces to calculate $W_{ox:tg}$ [25, 26], and constant $C_{ox:tg}$. For the highest accuracy, however, $C_{ox:tg}$ can not be assumed constant, according to the previous section:

- $C_{ox:tg} = C_p + C_f + C'_f$ for the inner and outer turn of the inductor.

- $C_{ox:tg} = C_p + 2C'_f$ for the other turns if more than 2 turns.

- $C_{ox:tg} = C_p + 2C_f$ for one turn inductors.

Fig. 11.1(b) illustrates the lumped inductor π-model and its capacitances for the spiral inductor with one terminal grounded. According to (11.19d), $K_{spi:tg} = 3$.

$$W = \frac{CV_o^2}{2} \tag{11.19a}$$

$$W_{ox:tg} = \int_0^L \frac{C_{ox:tg}}{2} \left(V_o - V_o\frac{x}{L}\right)^2 dx = \frac{C_{ox:tg}LV_o^2}{2}\frac{1}{3} \tag{11.19b}$$

$$W_{spi:tg} = \frac{C_{spi:tg}V_o^2}{2} = W_{ox:tg} \tag{11.19c}$$

$$C_{spi:tg} = \frac{C_{ox:tg}L}{3} \tag{11.19d}$$

Differentially Driven Symmetrical Center Tapped Inductor

The differentially driven inductor always has a zero signal potential point at the trace's center. It is sometime referred to as the center tap of the inductor and it can be connected to a DC potential for biasing purposes without damaging the inductor's properties. Since the potential is zero in the center, the symmetrical inductor can be seen as two spiral inductors with half the length and one terminal grounded. The capacitance $C_{sym:tg}$ is thus (11.20), [27].

$$C_{sym:tg} = \frac{C_{ox:tg}}{3}\frac{L}{2} = \frac{C_{ox:tg}L}{6} \tag{11.20}$$

In the ideal case without the resistive losses in Fig. 11.1(a) the differential capacitance is $\frac{C_{ox:tg}L}{12}$. According to (11.20), $K_{sym:tg} = 6$.

1.6   Energy Storage from Trace to Trace

Although the derivation of the concise expressions is complex, some simple approximate equations result. For most symmetrical geometries $K_{sym:tt}$ can be approximated by 2, and for spirals ($K_{spi:tt}$) by $1/n^2$, where n is the number of turns. Spiral inductors composed by many turns will therefore have a negligible trace to trace capacitance, as opposed to the symmetrical inductors where this capacitance has a large influence.

2.   The Inductor CAD-Tool Indentro

This section presents an optimization program for on-chip inductors based on the theory of the present chapter. By using empirical and semi-empirical formulas to calculate inductance, resistance, and capacitance, optimization time is greatly reduced, as compared to field solvers. The program also facilitates the drawing of a layout by generating an output file in cif format that can easily be imported into Cadence Virtuoso. Above all it incorporates the equivalent distributed capacitance analysis described in [7].
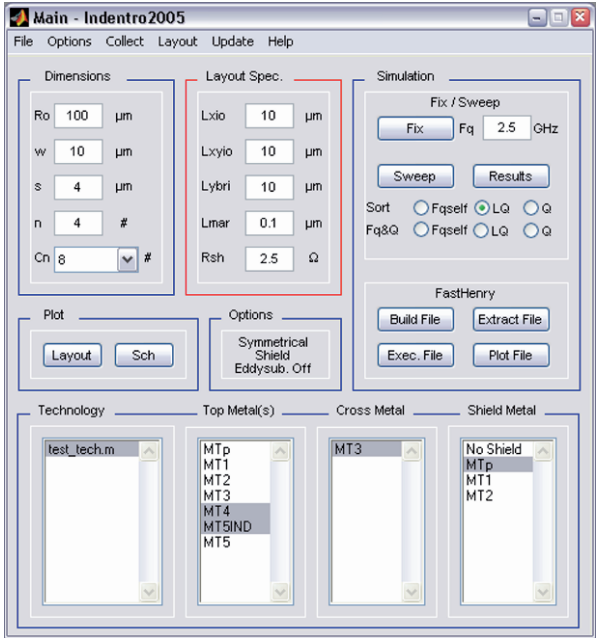
Figure 11.9. Main window of Indentro.

Indentro is a stand-alone Matlab program working under Windows, i.e., Matlab itself is not needed to run the program. Indentro can optimize the geometry of two of the most commonly used inductors, spiral and symmetrical, in terms of Q factor, LQ product, and self-resonance frequency (Fqself), as well as to promote the research carried out in [7]. Fig. 11.1 shows the currently supported geometries. It was to accommodate the increased demand for a fast and simple optimization program for monolithic inductors that Indentro was created, Fig. 11.9 [4].

Thanks to the simplicity of the algorithm, and using fast empirical and semi-empirical formulas to calculate the lumped π-model, Indentro reduces the optimization time greatly compared to field solvers. When a suitable geometry has been found using the fast formulas, an accurate π-model is created using FastHenry for the inductive and resistive parts, and the equivalent distributed capacitances from Indentro. This final equivalent lumped π-model is valid for a single frequency only, and when sweeping the frequency a new π-model is therefore calculated for each frequency.

2.1 Higher Accuracy Together with FastHenry

Since the concise and rapid empirical and semi-empirical formulas used by Indentro do not take into account the proximity effect, there can be a significant
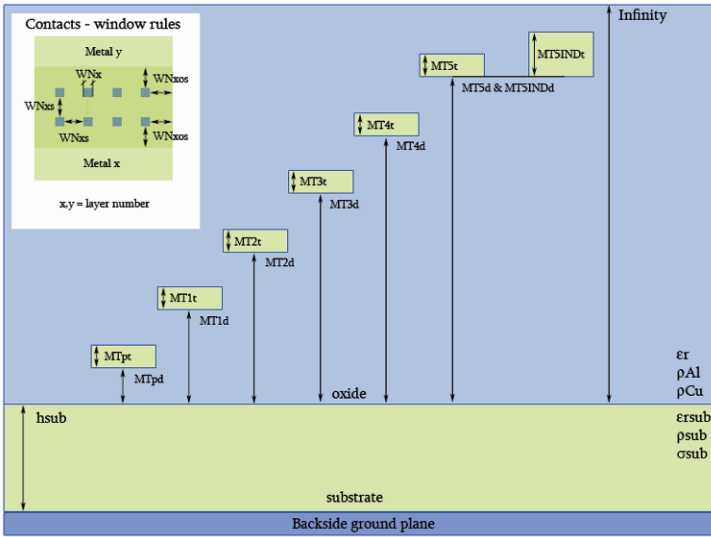
Figure 11.10.   Graphic view of technology file.

error in Q value. For this reason, Indentro can call up FastHenry and extract both the inductance and the series resistance [28, 9], while using capacitances by Indentro. The results from FastHenry may differ from the quicker formulas used by Indentro because of the proximity effect and because the final inductor geometry exported to FastHenry may not be exactly the same shape. The fast empirical formulas used by Indentro assume a certain geometry, however, after adding bridges and contacts, the geometry of the inductor may be more oval shaped in accordance with geometric parameters and layout rules. Thus, a FastHenry simulation is always needed when a particular inductor geometry has been determined. FastHenry does not take into account the electrical or magnetic losses of the substrate. When using a lightly doped substrate, shielded by a patterned ground-shield, the results will, however, be quite accurate.

2.2   Technology File

A technology file is needed containing all layer parameters of the semiconductor process. It is easy to create such a file by copying the "test_tech.m" file that is provided in the Indentro package [4] and changing or adding values. A graphic view of a technology file is shown in Fig. 11.10.

Export Layout to Cadence

A ".cif" file of the layout can be generated and imported into Cadence. To do this, the GDSII layer numbers of the drawing layers must also be known. This feature of the program is one of the most beneficial ones, since drawing the inductor traces (as well as contact windows) by hand is especially tedious work.

3. Verification of the Capacitance Model

The capacitance modeling described in this chapter and used by Indentro has been verified by on-chip probe measurements of different inductors in a $0.18\mu m$ CMOS technology [29], and by measurements of a quadrature oscillator in a $0.25\mu m$ CMOS technology [7], where the inductors have been designed so that the self-resonance frequency has a major influence on the oscillation frequency. In both of the cases the measurements and simulations agreed well.

3.1 $C_{ox:tt}$ For Naked versus Molded Die

It is assumed that the isolation material on top of the metal has infinite thickness. This is a fairly good assumption if the die is placed in a molded plastic package. However, with the die naked as in the case of probing this is not true, since the passivation material is usually less than $1\mu m$ thick. The capacitance, $C_{ox:tt}$, thus changes significantly. A change of $C_{ox:tt}$ is particularly hazardous for symmetrical inductors. The influence of this capacitance is large due to the low division factor of 2, compared to spiral inductors which have a factor close to the number of turns squared, n^2, see section 1.6. As can be seen in Fig. 11.6, for a very thin passivation layer, and $h_{package} = 0$, the field lines C_{gdt} travel mostly through air with $\epsilon_{air} = 1$ rather than SiO_2 dielectric with $\epsilon_r \approx 4$.

We now evaluate the difference in $C_{ox:tt}$ between a naked die and a molded one. The total capacitance between the strips are:

$$C_{ox:tt} = C_{gdt} + C_{gt} + C_{gdb} \tag{11.21}$$

Unless the strips are thick and narrow, and has a small spacing, C_{gdt} and C_{gdb} dominate over C_{gt}. Thus for strip lines with a thin passivation layer, and under the assumption that $C_{gdt} \approx C_{gdb}$ we get:

$$C_{ox:tt} \approx C_{gdt:mold}\frac{\epsilon_{air}}{\epsilon_r} + C_{gdb} \approx C_{gdt:mold}\left(1 + \frac{\epsilon_{air}}{\epsilon_r}\right) \tag{11.22}$$

Compared to $C_{ox:tt} \approx 2C_{gdt:mold}$ the capacitance $C_{ox:tt}$ decreases by a factor of (11.23) for a naked die compared to a molded package that Indentro assumes.

$$C_{sym:tt:die} = C_{sym:tt}\frac{1 + \frac{\epsilon_{air}}{\epsilon_r}}{2} \approx C_{sym:tt}\frac{1}{1.6} \tag{11.23}$$

Table 11.4.   Octagonal Inductor Geometries Under Measurement.

Inductor	Radius	Width	Spacing	Turns
No. I	$140\mu m$	$10\mu m$	$4\mu m$	8
No. II	$100\mu m$	$4.5\mu m$	$8.5\mu m$	6
No. III	$90\mu m$	$3\mu m$	$8\mu m$	6

The change of value of $C_{sym:tt:die}$ will be used later on when comparing the results of Indentro and FastHenry to the measurements. The change is less for inductors with thick metal and small spacing, when C_{gt} has more influence.

3.2   Probe Measurements

Three symmetrical inductors on four different chips have been measured. Two chips were measured with a 1-port method, and the other two with a 2-port. A die photo of two of the inductors can be seen in Fig. 11.11, and their geometries are presented in Table 11.4.

The inductors are not optimized for high Q-value, but in order to verify the results of Indentro they have rather been chosen with different parameter settings in terms of radius, width, spacing, and number of turns.

A $0.18\mu m$ CMOS process with 6 metal layers and high substrate resistivity ($10\Omega\text{-}cm$) was used, including a thick top metal option for inductors and power lines. The $2\mu m$ thick top metal of aluminum is located approximately $5\mu m$ above the metal 1 shield of the test inductors.

The measurements have been conducted with a HP8720C Vector Network Analyzer (VNA), which measures from 50MHz to 20GHz. Infinity probes from Cascade Microtech[1] was used. The probe measurements seemed to be sensitive to skating distances, and efforts were therefore made to maintain an equal skating length of $25\mu m$ for reproduceability of the results.

The VNA and probes were calibrated with an Impedance Standard Substrate (ISS), which provides open, short, load, and through structures. WinCal from Cascade Microtech was used to calibrate the VNA and collect the data. De-embedding of the measured data were performed using an open pad structure on-chip and simulated short structures with FastHenry (one for 1-port and another for 2-port).

1-Port Measurement

A balun was used to stimulate the inductors with true differential signals. The measurements were limited to the balun's frequency range, from 2GHz to 18GHz. To get the best differential signal from the balun, three different cables were tested on different outputs of the balun, thus a total of six combinations

(a) Inductor No. I

(b) Inductor No. III
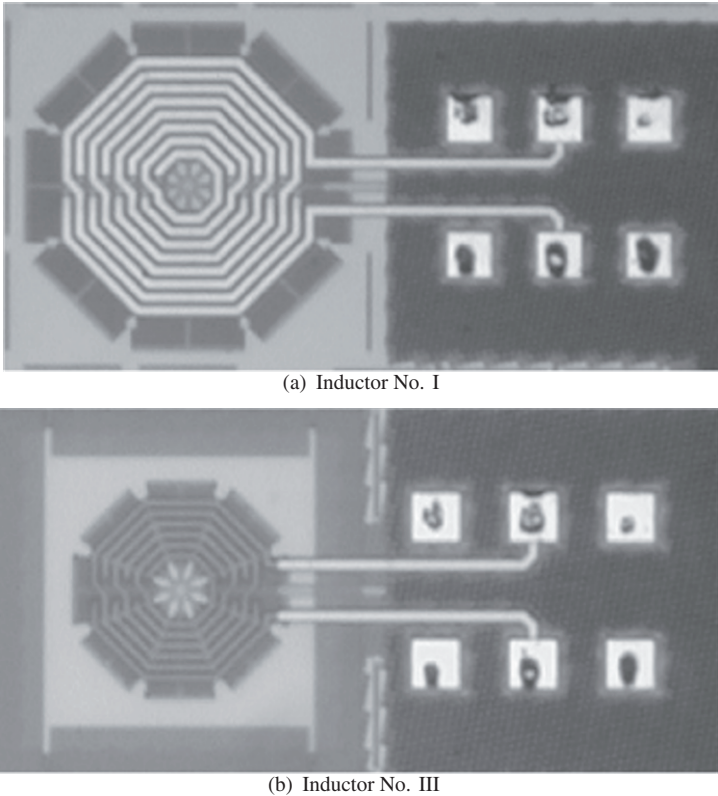
Figure 11.11. A photograph of two of the inductors. The skating marks from the probes are clearly visible.

were tested. The combination with the least phase error was the chosen and the 1-port measurements were compensated for the remaining error in phase and magnitude before extracting the differential Q-value.

Results

The differential Q-value and self-resonance frequency, f_{sr}, were extracted from the three inductors described in Table 11.4 and the results are plotted in Fig. 11.12. The Q-value was calculated by:

$$Q = -\frac{\Im m(Y_{in})}{\Re e(Y_{in})} \qquad (11.24)$$

where we have used $Y_{in} = Y_{11}$ for 1-port measurements, and $Y_{in} = Y_{11} - Y_{12}$ for 2-port. The self-resonance frequency can be found as the frequency where Q equals zero (13GHz for inductor II). As can be seen, the 1-port and 2-port measurements are in alignment with the simulation results by Indentro and

(a) Inductor No. I

(b) Inductor No. II

(c) Inductor No. III
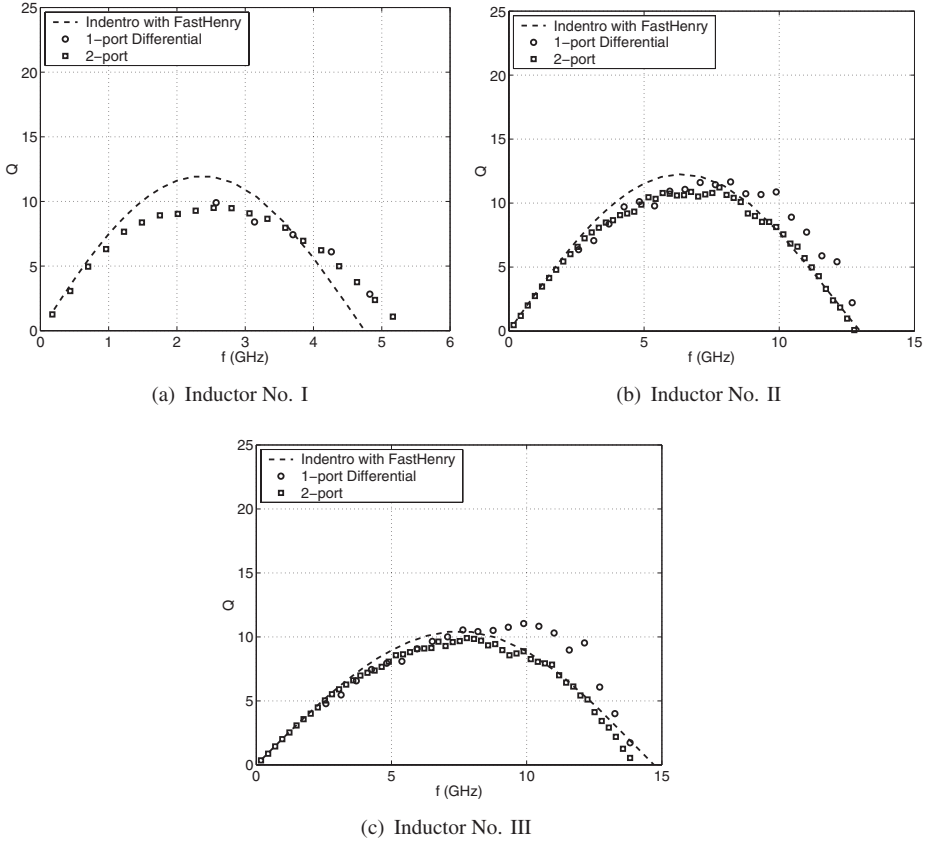
Figure 11.12. Comparison between Indentro with FastHenry, and 1-port and 2-port measurements.

FastHenry. Since the die is naked and not molded the capacitance $C_{sym:tt}$, calculated by Indentro, has been compensated with (11.23) for a better agreement in terms of Q-value and self-resonance frequency.

3.3   QVCO Frequency Measurements

A 1.8GHz QVCO with high phase noise performance suitable for low supply voltages has been implemented in a $0.25\mu m$ CMOS process from Agere Systems. Inductors were used instead of transistor based current sources to enable low supply voltages, and varactors in the buffer maximizes the output amplitude over the frequency range and also reduces the power consumption [7].

The circuits were packaged in Amkor $5 \times 5mm$ packages with 28 leads suited for RF applications, and mounted on a double sided PCB with SMA
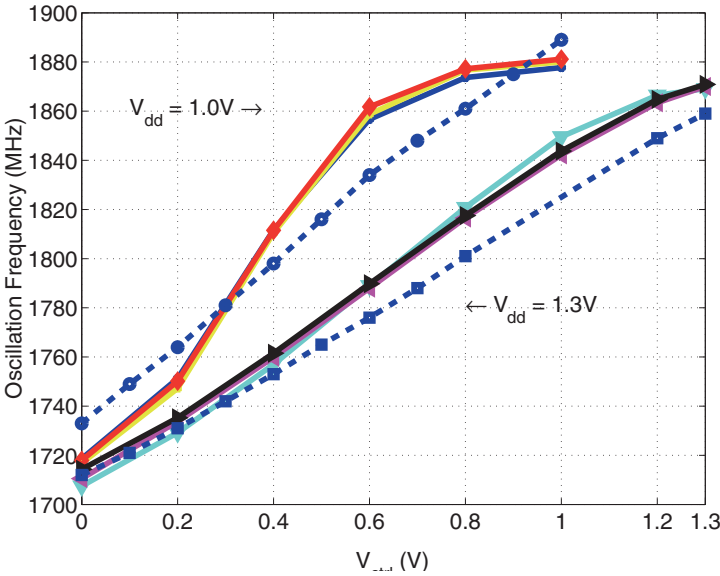
Figure 11.13. Tuning characteristic for three measured chips at 1.0V and 1.3V supply. Dashed lines is the simulated tuning characteristic.

connectors. Two supply voltages were tested, 1V at which the quadrature VCO and buffer used 2.5mA, and 1.3V at which the total current needed was 5.4mA.

Figure 11.13 presents the tuning characteristic for three chips measured at 1.0V and 1.3V supply. As can be seen the characteristic for the 1.3V supply is more desirable since it is almost linear. The dashed lines are the simulated frequency response. Fig. 11.14 illustrates the strength of the capacitance analysis presented, where frequency simulations have been made with and without the fringe capacitances. In this circuit the inductor capacitances represent approximately half the total tank capacitance, and the errors neglecting fringing becomes dramatic.

Notes

1 The probes are in a GSG configuration with a pitch of 100 μm. The frequency range is from DC to 40GHz. The contact resistance is less than 50 mΩ on aluminum pads.
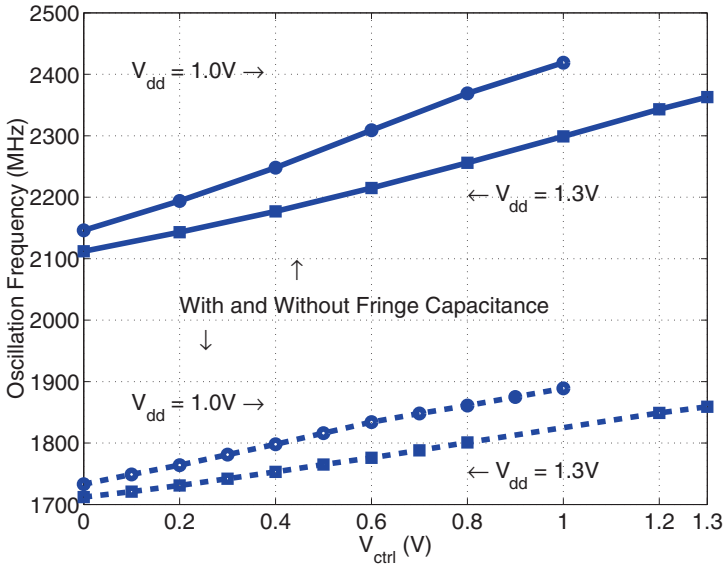
Figure 11.14.   Simulated tuning characteristic at 1.0V and 1.3V supply with and without the fringe capacitances.

References

[1] N. Troedsson and H. Sjöland, "High Performance 1V 2.4GHz CMOS VCO," in: *Proceedings 3rd Asian-Pacific Conference on ASICs*, pp. 185–188, August 2002, Taipei, Taiwan.

[2] E. Hegazi, H. Sjöland, and A. A. Abidi, "Filtering Technique to Lower LC-Oscillator Phase Noise," *IEEE J. Solid-State Circuits*, vol. 36, pp. 1921–1930, December 2001.

[3] F. Tillman, N. Troedsson, and H. Sjöland, "A 1.2 Volt 1.8 GHz CMOS Quadrature Front-End," *IEEE Symposia on VLSI Circuits*, 2004.

[4] N. Troedsson, "Indentro," http://www.indentro.com.

[5] K. T. Christensen and A. Jørgensen, "Easy Simulation and Design of On-chip Inductors in Standard CMOS Processes," in: *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 4, pp. 360–364, May 1998.

[6] C. P. Yue and S. S. Wong, "Design Strategy of On-Chip Inductors for Highly Integrated RF Systems (Invited)," in: *Proceedings of Design Automation Conference (DAC)*, pp. 982–987, June 1999.

[7] N. Troedsson and H. Sjöland, "A Distributed Capacitance Analysis of Co-Planar Inductors for a CMOS QVCO with Varactor Tuned Buffer Stage," *Analog Integrated Circuits and Signal Processing (AICASP)*, vol. 42 (1), pp. 7–19, January 2005.

[8]   A. M. Niknejad, "ASITIC," http://rfic.eecs.berkeley.edu/ niknejad/asitic.html.

[9]   M. Kamon, L. M. Silveira, C. Smithhisler, and J. White, *FastHenry User's Guide*, Boston: MIT Press, 1996.

[10] A. M. Niknejad and R. G. Meyer, "Analysis of Eddy-Current Losses Over Conductive Substrates with Applications to Monolithic Inductors and Transformers," *IEEE Trans. Microwave Theory Tech.*, vol. 49, no. 1, pp. 166–176, January 2001.

[11] S. R. Kythakyapuzha and W. B. Kuhn, "Modeling of Inductors and Transformers," in: *Proceedings of Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 283–286, May 2001.

[12] J.-H. Gau, S. Sang, R.-T. Wu, F.-J Shen, H.-H. Chen, A. Chen, and J. Ko, "Novel Fully Symmetrical Inductor," *IEEE Electron Device Lett.*, vol. 25, no. 9, pp. 608–609, September 2004.

[13] L. F. Tiemeijer, D. Leenaerts, N. Pavlovic, and R. J. Havens, "Record Q spiral Inductors in Standard CMOS," in: *Tech. Dig. of IEEE International Electron Devices Meeting (IEDM)*, pp. 949–951, 2001.

[14] S. S. Mohan, M. del Mar Hershenson, S. P. Boyd, and T. H. Lee, "Simple Accurate Expressions for Planar Spiral Inductances," *IEEE J. Solid-State Circuits*, vol. 34, no. 10, pp. 1419–1424, October 1999.

[15] H. A. Wheeler, "Simple Inductance Formulas for Radio Coils," in: *Proceedings of the Institute of Radio Engineers (IRE)*, vol. 16, no. 10, pp. 1398–1400, October 1928.

[16] A. Lofti and F. Lee, "Two Dimensional Skin Effect in Power Foils for High-Frequency Applications," *IEEE Trans. Magn.*, vol. 31, pp. 1003–1006, March 1995.

[17] J. N. Burghartz and B. Rejaei, "On the Design of RF Spiral Inductors on Silicon (Invited)," *IEEE Trans. Electron Devices*, vol. 50, no. 3, pp. 718–729, March 2003.

[18] Y. E. Chen, D. Bien, D. Heo, and J. Laskar, "Q-Enhancement of Spiral Inductor with N^+-Diffusion Patterned Ground Shields," *IEEE MTT-S International Microwave Symposium Digest*, vol. 2, pp. 1289–1292, May 2001.

[19] J. N. Burghartz and B. Rejaei, "Effects of Dummy Patterns and Substrate on Spiral Inductors for Sub-micron RF ICs," in: *Proceedings of Radio Frequency Integrated Circuits (RFIC) Symposium*, pp. 419–422, June 2002.

[20] D. Kehrer, "Design of Monolithic Integrated Lumped Transformers in Silicon-based Technologies up to 20 GHz," Diplomarbeit, Vienna University of Technology, December 2000.

[21] T. C. Edward, *Foundations for Microstrip Circuit Design*, John Wiley, New York, 1981.

[22] K. C. Gupta, R. Garg, I. Bahl, and P. Bhartia, *Microstrip Lines and Slotlines*, Artech House, Norwood, second edition, 1996.

[23] Z. Jiang, P. S. Excell, and Z. M. Hejazi, "Calculation of Distributed Capacitances of Spiral resonators," *IEEE Trans. Microwave Theory Tech.*, vol. 45, no. 1, pp. 139–142, January 1997.

[24] J. Sieiro, J. M. López-Villegas, J. Cabanillas, J. A. Osorio, and J. Samitier, "A Physical Frequency-Dependent Compact Model for RF Integrated Inductors," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 1, pp. 384–392, January 2002.

[25] C.-H. Wu, C.-C. Tang, and S.-I. Liu, "Analysis of On-Chip Spiral Inductors Using the Distributed Capacitance Model," in: *Proceedings of Asian-Pacific Conference on ASICs*, 2002.

[26] A. Zolfaghari, A. Chan, and B. Razavi, "Stacked inductors and transformers in CMOS technology," *IEEE J. Solid-State Circuits*, vol. 36, no. 4, pp. 620–6282, April 2001.

[27] M. K. Mills, "Inductive Loop system Equivalent Circuit Model," in: *Proceedings of IEEE Vehicular Technology Conference (VTC)*, vol. 2, pp. 689–700, May 1989.

[28] "Fast field solvers," http://www.fastfieldsolvers.com.

[29] N. Troedsson, J. Wernehag, and H. Sjöland, "Measurements of Differential Symmetrical Inductors," in: *Proceedings of IEEE Norchip Conference*, 2005.

Chapter 12

CHALLENGES IN THE DESIGN OF PLLS IN DEEP-SUBMICRON TECHNOLOGY

Waleed Khalil and Bertan Bakkaloglu

1.   Introduction

Phase-locked loops (PLLs) are used to implement a variety of timing related functions, such as frequency synthesis and clock/data recovery. Previously PLL designers have been preoccupied with optimizing one or several PLL parameters to achieve the best performance. More recently however, the challenge has shifted to building affordable radios by seamless integration of different circuit blocks. This includes building the PLL block in deep-submicron process nodes that are hostile to pure RF and analog circuit techniques. In this chapter different challenges in the design of PLLs in deep submicron technology to address the growing need for developing multi-band radios and realizing high data rate communication systems will be explored. Furthermore, some of the digital techniques and architectures for designing PLLs that are now becoming more ubiquitous in digital centric process technologies will be examined.

2.   Technology Trends

The recent advances in deep submicron technology have enabled the economic integration of all the necessary ingredients to build multiple radio systems on a small Si area. However, with simultaneous operation between some or all of these radios, the challenge remains to achieve a noise floor that is in par with single radio chip solutions. While interference noise (substrate coupling, inductance and signal integrity, voltage headroom shrinkage) remains to be solved, other challenges have emerged and need to be addressed. As process dimensions shrink, gate current and subthreshold leakage present new challenges to
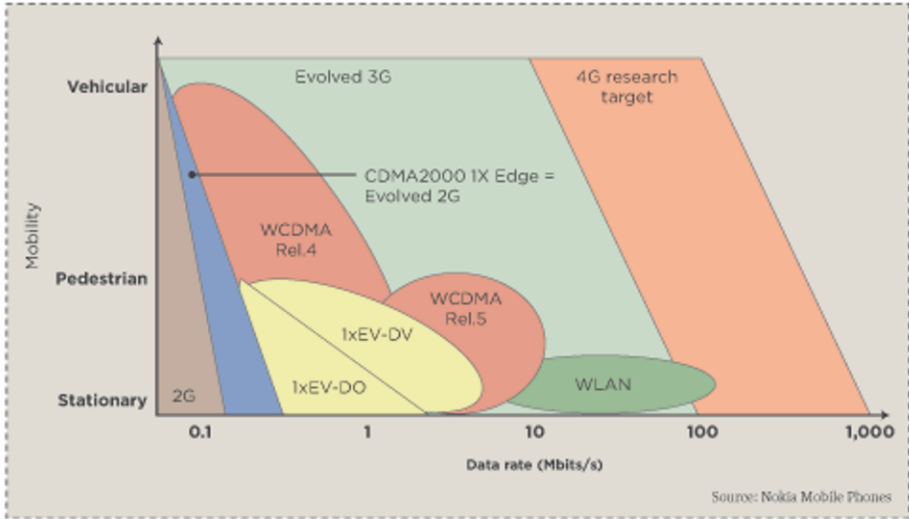
Mobility and data rate for radio systems.

Figure 12.1.   Data rates for current and future radios.

both system and circuit designers. On the system front, leakage current is becoming a major factor in system power budgeting and thus has a large impact on all battery-powered devices. On the circuit side, leakage is challenging the traditional charge based circuit design techniques. Technology scaling is also having an ever-increasing effect on process variations. Device mismatch and threshold voltage variation are now impacting the design in many different ways. A new paradigm shift is required, where statistical design methodology will replace traditional and conservative "over-design" methods.

2.1   Multi-Standard/Multi-Band Design

The number of mobile subscribers has grown more than a hundredfold in the past 10 years with over 1 billion subscribers globally. A look at the evolution of radio systems in wide area networks (WAN) from the current second-generation (2G and 2.5G) to third-generation (3G) reveals the rapid convergence between voice and data systems. With respect to local area networking (LAN), the rapid acceptance of WLAN and Bluetooth technologies has increased the rate of their attachment to mobile terminals (e.g. mobile PCs and cellular phones), as highlighted in Figure 1. New technologies such as HSDPA (High Speed

Downlink Packet Access), EVDO (Evolution Data Optimized) and WiMAX are being deployed and adopted at a never before seen pace. GPS and DVB-H (digital video broadcast) are also among other technologies that are leaping into wireless terminals. Figure 1 highlights the increased demand for data rate in both current and future wireless standards. The idea of integrating multiple radios in a system on a chip (SoC) or in a package (SiP) may have been ridiculed a few years ago- but is now being touted as a viable option. It has only been a few years since the mobile terminal industry created a common platform where two or more radios were co-located on the same PCB board. The challenge at the time was to make this integration happen in a cost effective way without incurring significant cost. Today, the focus is more on the total integration of multiple radios on one piece of silicon with the same goal of making future wireless products more affordable. In order to make this a reality, there are numerous obstacles that need to be overcome in the RF, baseband, front-end module, packaging, CAD and system validation fronts. In the RF domain, coexistence (i.e. radio to radio interference) is a major issue that needs to be carefully analyzed and addressed. A closer look at the multi-radio frequency spectrum (Figure 2) shows that both RF circuit and system engineers are confronted with multiple interference problems ranging from band coexistence (Intermodulation, receiver saturation) to out-of-band emissions. The required signal isolation at the package or substrate crosstalk level can be upward of 100dB. Previously, it was possible to verify this level of isolation by performing an EM (electromagnetic) simulation at the small scale single chip level. Unfortunately, with the given multi-radio complexity, RF designers are unequipped to simulate systems with such complexity due to the shortcomings of the existing EM tools.

With regards to the PLL, the multi-standard multi-band radio requirements present a new set of challenges and unresolved issues. Developing the appropriate usage model for a multi-radio device is a key factor in determining the number of PLLs required to operate in parallel and the specifications for each one of them. Table 1 provides a list of the different standards that will most likely be supported in a multi-radio solution. Given that it is still unclear which combination of these standards will be operated simultaneously, it is difficult to determine the required number of discrete PLLs on the IC. However, some logical assumptions can be made based on the predicted user model in a multi-band radio. For instance, either WiMAX or any one of the cellular standards (GSM or WCDMA) will need to be supported at any given time depending on the carrier and network availability. Concurrently, WiFi, Bluetooth, GPS and DVB-H could all be operating in parallel. Hence, in an SoC it is possible to have six separate PLLs that are operating simultaneously (WCDMA is FDD which requires two PLLs). Therefore, it is unavoidable to have mixing between the frequencies and associated harmonics of these different PLLs at the power sup-
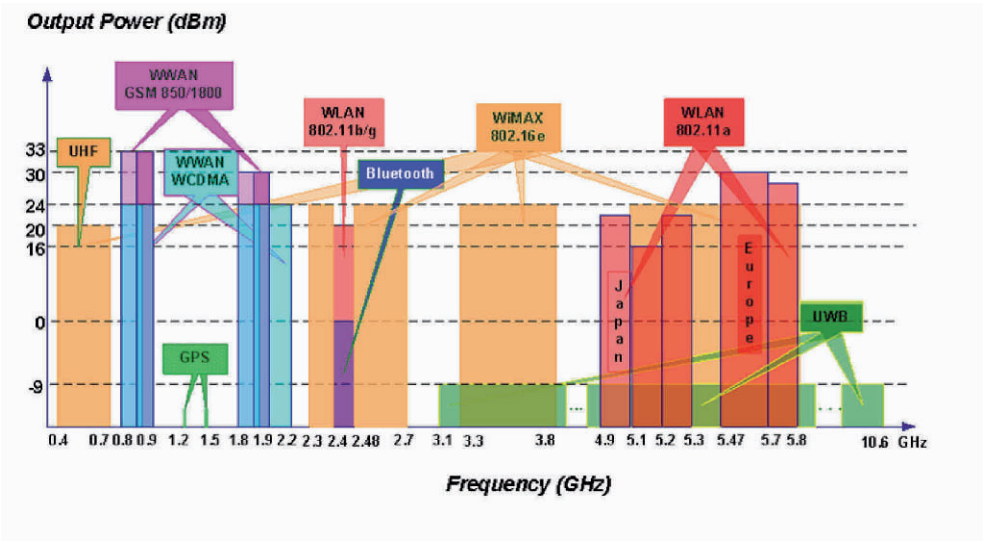
Figure 12.2. Mobile frequency spectrum

ply or substrate level. Unless this mixing is carefully analyzed and prevented, it could lead to undesired spurs or even a complete loss of lock in one or more of the PLLs. In an ideal scenario, all of these separate PLLs are built based on a single PLL architecture that can be configured for the different frequency bands and standard requirements. Although this will minimize the design cycle at the chip-top assembly level, it will make the design of the individual PLL extremely challenging if not impossible: First, this will require a VCO design with a very wide tuning range that will cover over a decade of frequency bands. This can be done using: 1) a very large varactor which translates to large a VCO gain (e.g. few GHz/V) and thus a poor phase noise performance or 2) a switched capacitor array which has never been proven to cover a tuning range of more than 25%. Secondly, using a single loop filter design to cover all the different standards is not possible as a tradeoff between the PLL switching time and close-in/far-out phase noise needs to be made per each standard specifications. It is then clear that building a unified PLL or frequency synthesizer using the traditional analog approach will not be possible and hence a more digital centric approach will be needed. Some of these new options will be further discussed in the following sections.

2.2   Coexistence with Digital Nanometer Technology

Digital CMOS is clearly the process of choice for multi-standard multi-band radio integration. This technology has already been proven with respect to technological and commercial issues, resulting in today's wide availability of

Table 12.1. Multi-radio PLL frequencies and bandwidth Standard RX Frequency Range (MHz) TX Frequency Range.

Standard	RX Frequency Range (MHz)	TX Frequency Range (MHz)	BW (MHz)
GSM:			
850	824-849	869-894	
900(E-GSM)	880-915	925-960	0.2
1800	1710-1785	1805-1880	
1900	1850-1910	1930-1990	
WCDMA:			
Band I-FDD	1920-1980	2110-2170	
Band II-TDD	1850-1910	1930-1990	5
Band III-TDD	1710-1785	1805-1880	
Band V-TDD	824-849	869-894	
WiFi:			
802.11b/g		2400-2480	20
802.11a		4900-5850	
Bluetooth		2402-2480	1
WiMAX:			1.25,1.75,2.5,
2.5 GHz(Licensed)		2300-2700	3.5,5,7,10,
3.5 GHz(Licensed)		3300-3900	14,20
5.5 GHz(Unlicensed)		5150-5850	
GPS		1575.42	N/A
DVB-H:			
UHF		470-890	5,6,7,8
L-Band		1670-1675	

single chip radio solutions. Moore's law of scaling and the efficient implementation of DSP functions with RF and mixed-signal circuits on a single chip are what make this technology superior compared with other technologies. Digital CMOS technology however, has several disadvantages when compared with the widely available BiCMOS technology. Issues such as high leakage and 1/f noise, reduced headroom, large coupling due to substrate or transistor gate, inferior passive elements and poor RF models still need to be addressed. The impact of technology scaling on the device performance is shown in table 2. While it is clear that the transistor speed (i.e. f_T) continues to scale up as features scale down, it is also evident that this gain is made at the expense of large loss of other performance parameters. For instance, there is an exponential rise in sub-threshold and gate leakage current. Also, the analog performance of the transistor represented by its gmro has greatly suffered with process shrink.

With the advance of technology nodes, the growing challenge of DFM (design for manufacturing) presents another problem for RF circuit designers. As 180nm designs shrinks to 130nm, 90nm, 65nm, and smaller, the shift to sub-wavelength lithography and high-K dielectrics requires a new paradigm for manufacturing robust circuits. With respect to design, this implies a rapid in-

Table 12.2.   CMOS technology scaling trends (C. Sodini, RFIC2005).

	0.25 um		0.18 um		0.13 um		0.09 um	
V_{dd} (V)	2.5	1x	1.8	0.7x	1.2	0.5x	1.0	0.4x
I_{on} (uA/um)	600	1x	600	1x	550	0.9x	850	1.4x
I_{off} (uA/um)	0.01	1x	0.02	2x	0.32	32x	7	700x
I_{on}/I_{off} (10^6)	60	1x	30	0.5x	1.7	0.03x	0.12	0.002x
I_{gate} (nA/um)	2e-5	1x	0.002	100x	0.65	3e+4x	6.3	3e+5x
g_m (mS/um)	0.3	1x	0.4	1.3x	0.6	2x	1.0	3.3x
g_o (uS/um)	7.7	1x	15	2x	42	5.4x	100	13x
$g_m r_o$	39	1x	27	0.7x	14	0.36x	10	0.26x
f_T (GHz)	30	1x	60	2x	80	2.7x	140	4.7x

crease in the number of design rules (DRs), which greatly restricts that degrees of freedom needed to optimize both the circuit and layout performance. In general, the trend in DFM is towards: 1) using only a limited set of device geometries, and 2) enforcing a tight poly, diffusion, metal and via density by adopting the auto dummification process across the entire Si die. Although this process of dummification can be easily applied in the digital world, it is an extremely challenging process when it comes to analog or RF designs. Unlike digital layout, both analog and RF layouts are structured for matching and signal isolation purposes where it usually takes many rounds of iterations to deal with coupling and device/signal matching issues. Having a dummification script inserting random structures over the whole layout to satisfy a given density requirement destroys what took a long time for RF and analog designers to build. With respect to the PLL, the VCO and charge pump are the ones that are greatly affected by the dummification process. In a VCO layout, the tank inductor occupies a large area (e.g. $2000\mu m \times 200\mu m$) and is typically drawn with a gap of a few tens of microns to the nearest active or metal area. When dummification is randomly applied inside and around the tank area it creates a large disturbance to the magnetic flux. This in turns makes the inductance and quality factor much more difficult to predict and control. The PLL LPF occupies a large die area (e.g. few mm^2) and is typically implemented with either MIM or MOS caps. Having a large area of the die filled with a uniform structure (cap) breaks the dummification rules and the only alternative is to spread the filter cap across the entire die. Inside the PLL, the filter cap is tied to the VCO control voltage, which is the most sensitive area in the PLL design. Hence, the chance of large noise coupling to this node will increase tremendously causing a large amount of spurs to appear at the PLL output. In general, these kinds of
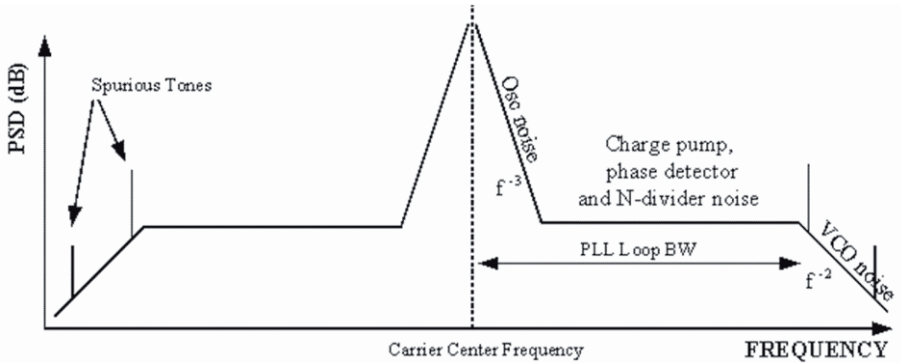
Figure 12.3.   Power spectral density of a PLL output signal.

spurs are difficult to debug and eliminate. It is clear from the above discussion that technology scaling will make the task of designing a multi-band radio too difficult unless new circuits and architectures are invented to fully exploit the speed of digital technology. It is then natural to assume that future PLL architectures will be based on over-sampling techniques accompanied with fast DSP operations.

3.   PLL Performance Metrics

In this section, a review of the basic PLL performance parameters will be presented. Although the majority of these parameters have a direct influence on certain standard specifications, they are rarely specified in a product datasheet and instead are calculated based on detailed system simulations.

3.1   Spectral Emission

Spectral emission is a measure of the PLL output power versus offset frequency in a given band of interest. In general, there are two different types of phase instabilities in a PLL output signal, as shown in Figure 3. The first is deterministic -commonly called spurious- which is caused by discrete signal frequencies appearing as discrete components in the power spectral density (PSD) plot. Typically, this is caused by coupling through either the power line or by direct signal coupling due to capacitive or inductive paths into the VCO nodes. The second type of phase modulation is random in nature and is termed phase noise.

In modern communication systems, signal sources are used in both modulation and demodulation processes. With respect to modulation, a local oscillator (LO) signal is used as a carrier signal in order to up-convert the baseband information. Conversely, an LO signal is also required to achieve frequency

down-conversion in the demodulation process. With the rapid and ubiquitous usage of available spectrum, RF system designers are forced to adopt stringent specifications for the allowable frequency tolerance of the signal sources. In a simple case, the spectrum of a spectrally pure, single frequency sinusoidal source is represented as an infinitely narrow pulse (i.e. Dirac-delta function) in the frequency domain. In reality, the spectrum of a real carrier signal has finite width as it suffers from undesired amplitude and phase modulation due to noise disturbances. Digital communication systems specify the Bit Error Rate (BER) as a direct measurement of the link quality. The degradation to the BER is directly proportional to the amount of unwanted amplitude and phase modulation to the signal. In practice, spurious phase modulation is more detrimental than spurious amplitude modulation. This is due to the fact that the vast majority of digital communication systems rely on angle (phase or frequency) modulation instead of amplitude modulation. Also, the magnitude of spurious amplitude error that the signal suffers is usually far less than spurious phase modulation and can also be easily calibrated unlike the phase modulation error.

Phase Noise & Spurs

Phase noise in an oscillator signal is defined as rapid, short-term, random fluctuations in the phase of a wave caused by time-domain instabilities. Phase noise shows as continuous sideband noise in the PSD and is caused by thermal, shot or flicker (1/f) noise sources. In the time domain, an actual sine wave signal with both amplitude and phase modulation could be represented as:

$$v(t) = V_o[1 + A(t)]\sin[2\pi f_0 t + \phi(t)] \qquad (12.1)$$

where $A(t)$ denotes the amplitude variation/modulation part and $\phi(t)$ denotes the phase fluctuations. There are two fundamental methods to characterize phase perturbations1. The first is done directly in the RF domain by measuring the PSD of the signal on a spectrum analyzer where phase noise is represented in the sideband power on either side of the carrier f_0. Phase noise is then extracted by measuring the spectral density of these sidebands at a given offset, $S_v(f_0 \pm f)$. The second method is by demodulating the carrier and then extracting the phase fluctuation term at baseband. The analysis of this baseband signal can produce the spectral density of phase fluctuations, $S_\phi(f)$, or the frequency fluctuations, $S_f(f)$. The relation between phase noise and $S_\phi(f)$, $S_f(f)$ will be detailed later in the discussion. In the following sections, each of the different PSD components will be analyzed.

1. $S_v(f_0 \pm f)$:

This the PSD of the voltage fluctuations measured directly using a spectrum analyzer in Watts/Hz. The sideband energy measured is usually comprised of both AM and PM noises. Assuming that AM noise is negligible compared to PM noise, $S_v(f_0 \pm f)$ is the most straightforward technique to measure the phase fluctuations of a signal. However, to extract the correct signal noise, the phase noise sidebands to be measured must be greater than the spectrum analyzer's own noise floor by at least 10 dB.

2. $S\phi(f)$:

This is defined as the spectral density of the rms value of the phase fluctuations measured in $radians^2/Hz$:

$$S_\phi(f) = \frac{[\phi_{RMS}(f)]^2}{df} \qquad 0 < f < \infty \qquad rad^2/Hz \qquad (12.2)$$

Note that this measure of the phase modulation sideband noise power (i.e. $\phi(t)$) is not related to the carrier frequency. In practice, $S_\phi(f)$ is measured in baseband by passing the signal through a mixer/phase detector and then measuring the PSD of the output signal. This is not to be confused with $S_v(f_0 \pm f)$ which is PSD of the signal itself measured in Watts/Hz. Equation (2) can be used to extract the rms phase error as follows:

$$\phi_{RMS}^2 = \int_0^\infty S_\phi(f)df \qquad (12.3)$$

3. $\mathcal{L}(f)$:

The single sideband phase noise, $\mathcal{L}(f)$, is defined as the noise power due to the phase fluctuations of the signal in a $1Hz$ bandwidth to the total carrier power, specified at a given offset fHz from the carrier. Although this is an indirect representation of phase noise, it is by far the most commonly used term to express phase noise. Assuming that AM noise is much less or further suppressed from the PM noise, $\mathcal{L}(f)$ can be directly extracted from the spectrum analyzer as the ratio of noise sideband power to the carrier power; as shown in Figure 4:

$$\mathcal{L}(f) = 10\log \frac{S_v(f_0 \pm f)}{P_{signal}} \qquad dBc/Hz \qquad (12.4)$$

However, care should be taken as noted earlier when using spectrum analyzers to measure $\mathcal{L}(f)$ as they generally have a relatively high noise level that limits the accuracy of phase noise measurement. Typical spectrum analyzers can reliably measure phase noise down to $-130dBc/Hz$ level. Another method to extract $\mathcal{L}(f)$ is by measuring $S_\phi(f)$ [11]. Recall that in basic FM modulation theory, the carrier sideband levels are related to the magnitude of
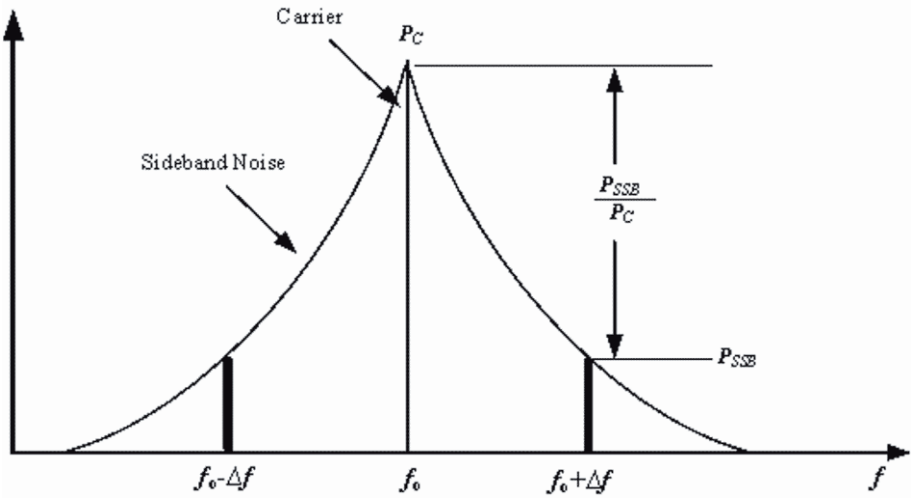
Figure 12.4.   Spectrum analyzer output displaying the carrier phase noise.

phase deviation by Bessel function terms. Assuming the phase deviations are sufficiently small to prevent high order sidebands, only the first Bessel terms are significant and other terms can be neglected. Then the relation between the single sided phase noise, $\mathcal{L}(f)$ and the spectral density of the phase fluctuations $S_\phi(f)$, can be a approximated by:

$$\mathcal{L}(f) \approx 10 \log \left(\frac{S_\phi(f)}{2} \right) \qquad dBc/Hz \qquad (12.5)$$

Hence, there is numerical equivalency between $dB(rad^2/Hz)$ and dBc/Hz. The relation between the rms phase error and phase noise can then be calculated by substituting Eq. (5) into Eq. (3):

$$\phi_{RMS}^2 = 2 \int_0^\infty 10^{\frac{\mathcal{L}(f)}{10}} df \qquad (12.6)$$

In general, there are two different types of phase and amplitude noise; either additive or multiplicative. Additive noise is defined when the noise power is independent of the signal power. The most common source of additive noise in a system is the amplifier added noise, $kTBF$; where B is the bandwidth of interest and F is the noise factor of the amplifier [3]. Assuming an amplifier with noise figure, $NF(dB)$, the total noise power per Hz referred to the input of the amplifier is:

$$N = KT + NF(dB) = -174dBm/Hz + NF(dB) \qquad (12.7)$$

where N is the total noise power per unit Hz at the input of the amplifier. The ratio between the total noise power to the input carrier level in dBc/Hz can be expressed as:

$$\frac{N}{C} = -174dBm/Hz + NF(dB) - C(dBm) \qquad dBc/Hz \quad (12.8)$$

where C is the input carrier power in dBm. The above equation includes both AM and PM noise contributions, where half of the total noise is AM noise and the other half is PM noise. The phase noise part of the signal can then be expressed as [11]:

$$\mathcal{L}(f) = \left(\frac{N}{C}\right)_\phi = \frac{N}{2C} = -177dBm/Hz + NF(dBm) - C(dBm) \quad (12.9)$$

Note that the above equation expresses the amplifier added phase noise to the signal assuming the amplifier is operated in the linear region. For the case where a high powered signal is present at the input of the amplifier due to either the carrier itself or a blocker signal, the amplifier will be operated near its compression point and the NF generally increases by a few dBs. Additive noise affects the PLL output signal as it gets buffered and amplified before being sent outside the PLL. In order to avoid loading the LC tank in the VCO with a variable load, an isolation buffer is usually added, where the NF of this buffer sets the wideband phase noise floor level according to Eq. (9). Among all multi-radio standards in Table 1, the GSM standard has the most stringent wideband noise requirement of $-162dBc/Hz$ at $20MHz$ carrier offset. In order for the VCO buffer to have minimal noise contribution on the output signal, it is typically designed to have phase noise not exceeding $-170dBc/Hz$. Assuming a VCO swing signal of $0dBm$ ($0.63V_{p-p}$), this corresponds to a worst case NF of $7dB$ for the VCO buffer. Fortunately, as technology scales down, the switching time of the transistors improves and meeting this noise figure target is no longer viewed as a design obstacle. Another type of noise that occurs when the noise power is proportional to the signal power is multiplicative noise [3]. This type of noise is associated with the presence of noise components in either the amplitude gain path (i.e. AM noise) or phase gain path (PM noise) of a given circuit. For the case of a VCO circuit, an amplifier is used as a gain stage to compensate for different losses in the oscillator tank. The thermal and 1/f noise components of the amplifier's transconductance (g_m) are converted to multiplicative AM and PM noises at the output of the oscillator. In addition to the PM noise part, the AM noise will also modulate the varactor element, which is part of the tank circuit, thus causing the output frequency to change and resulting in AM to PM noise conversion [1].

With technology scaling, the impact of this noise on the VCO low frequency noise can be significant unless special mitigation techniques are introduced in the VCO design. As process shrinks, the 1/f noise corner of the transistors has significantly increased to the $100 - 1000MHz$ range and thus causes this type of noise to impact the in-band, out-of-band and wideband noise spec of the PLL. The first step in reducing this noise is to use a very low gain VCO architecture; e.g. $K_v < 50MHz/V$. This can be achieved by using a switched cap array along with a small varactor to constitute the VCO tuning element. Another technique to reduce the AM to PM noise contribution is by using an amplitude limiting circuit to control the VCO bias, thus greatly minimizing the total amplitude variation at the VCO output. The analysis of both of these techniques will be further discussed in the following sections. Phase noise imposes fundamental limitations on the ultimate SNR, which can be achieved when detecting either an FM or a PM modulated signal. The majority of today's communications systems rely on phase modulation techniques (e.g. QPSK or 8PSK). For these systems, in order to meet the required SNR, the maximal allowable phase error and/or rms phase error are typically specified for both the transmitted signal and the receiver LO signal. Alternatively, for systems using more complex amplitude and phase modulation schemes (e.g. QAM), the Error Vector Magnitude (EVM) is typically specified. The EVM is a measure of both amplitude and phase errors as both are important in determining the overall SNR of the signal. Since phase error is not directly specified for these systems, it is then up to the system designer to allocate a certain budget in the EVM for phase error, leaving the rest to amplitude error. However, in a properly designed synthesizer, the amplitude error is rarely a consideration since it is usually $20dB$ or more below the phase error level. This is attributed to the fact that the major source for amplitude error (which is quadrature amplitude mismatch), can be easily reduced by either layout matching techniques or by on-chip calibration techniques. LO phase noise can degrade both the receiver's sensitivity and selectivity, as shown in Figure 5. Close-in (i.e. in-band) phase noise of the LO desensitizes the received signal by self mixing with the desired signal and thus causing a reduction in the SNR. Conversely, in a process termed reciprocal mixing, wide-band (i.e. far-out) phase noise of the LO mixes with interferers, resulting in additive broadband noise from the interferers.

A) Close-in Phase Noise

Close-in phase noise impacts the received signal as it gets demodulated along with the desired signal by causing adjacent constellation points to be incorrectly detected. Figure 6 shows the constellation diagram of a QPSK signal and the effect of phase noise on its demodulation. The original signal for bits (1,1) is shown as the dark circle while the signal effected by the LO phase noise is shown as the light circle. The phase error between the ideal signal and the actual received signal is statistically distributed and typically modeled using a
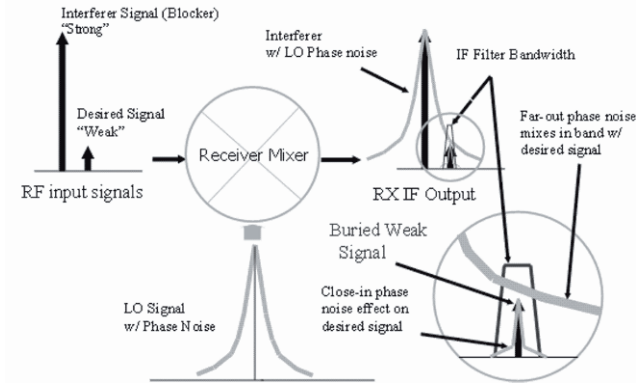
Figure 12.5. Impact of LO close-in and far-out phase noise on the receiver.

Gaussian distribution with the standard deviation equal to the rms phase error of the LO. Since the rms phase error represents only one standard deviation in the phase rotation, large rms phase errors would increase the probability in making an error. The extent of the system's vulnerability to phase error depends greatly on how closely packed the constellation points. Complex modulation schemes such as 16QAM and 64QAM have more compact constellation diagrams than simple BPSK or even QPSK constellations and are therefore much more susceptible to phase error.

The uncoded symbol error rate (SER) for any given modulation scheme can be computed as follows [2]:

$$P_{SER}\left(\frac{E_s}{N_0}, \phi\right) = \int_{-\pi}^{+\pi} P_s\left(\frac{E_s}{N_0}\Big|\phi\right) P_\phi(\phi, \sigma_\phi^2)d\phi \qquad (12.10)$$

Where E_s/N_0 is the ratio of the symbol energy to noise power spectral density. The first quantity inside the above integration $P_s(E_s/N_0|\phi)$, is the conditional probability of making a symbol error given a certain phase error ϕ. This quantity is dependent on the modulation type and is derived by Crawford [2] for different modulation schemes. The second quantity $P_\phi(\phi, \sigma_\phi^2)$ is the probability density function of phase error and can be calculated using the Tikhonov probability distribution function:

$$p_\phi(\phi, \sigma_\phi^2) = \sqrt{\frac{1}{2\pi\sigma_\phi^2}} \exp\left[\frac{\cos(\phi) - 1}{\sigma_\phi^2}\right] \qquad (12.11)$$

where σ_ϕ^2 is the variance of the phase error which is set to equal the rms phase error, $\phi_r ms^2$, as phase noise is a randomly distributed process with zero mean.
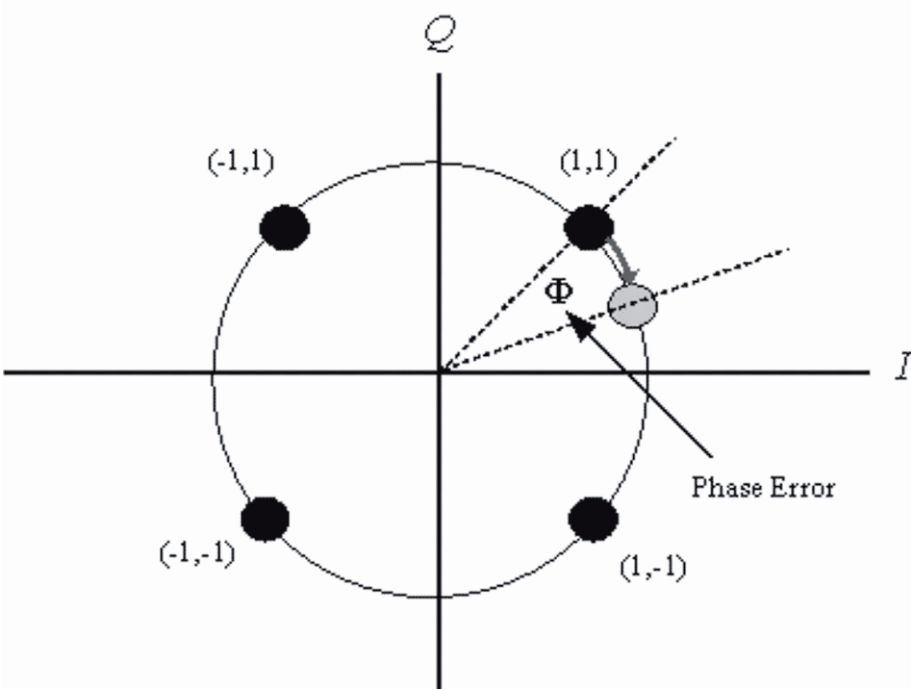
Figure 12.6. Effect of phase noise on QPSK signal.

Equations (10), (11) and (6) can be used to calculate the impact of a given PLL phase noise on the SER or BER for a given modulation type.

B) Wideband Phase Noise

The phenomenon of reciprocal mixing in the receiver is illustrated in Figure 5. The magnitude level of the noise sidebands has been exaggerated to highlight the problem [9]. In practice, since the interfering signal is generally many decades higher than the desired signal, the phase noise of the LO at large offset must be kept at a much lower magnitude to prevent further SNR degradation. For narrow band signals, the phase noise response in the region of interest can be assumed to be flat and the maximum allowable phase noise of an LO signal can be derived assuming a given SNR as:

$$\mathcal{L}(f) \leq P_{desired}(dBm) - P_{interferer}(dBm) - SNR(dB) - 10\log(B)$$

$$(12.12)$$

where B is the channel bandwidth. In reality, a margin of few dBs $(3-5dBs)$ is added to the above equation to prevent phase noise from becoming the dominant factor in limiting the receiver performance. With respect to the transmitter,
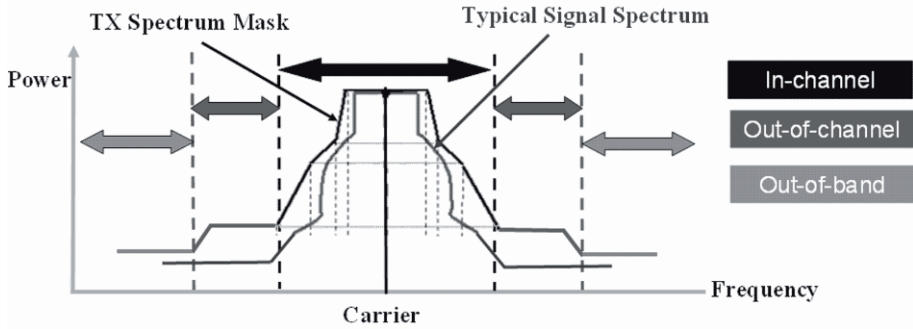
Figure 12.7. Typical transmitter frequency mask.

wideband noise results in the radiation of significant energy into one of the adjacent TX channels or to the nearby RX band. High transmitter power can also jam nearby receivers operating on the adjacent channel. A frequency mask is typically specified for the transmitted power in any given standard. The mask is a measure of the transmitter output power versus frequency and is strictly monitored by the standard body (i.e. FCC) to insure coexistence between multi-channels, multi-bands and multi-standards. There are three different emission regions that are specified in a transmitter mask: 1) in-channel, 2) out-of-channel and 3) out-of-band signal emission, as shown in Figure 7. In-channel emission determines the link quality by limiting the in-band spurious emission in the allocated channel. Out-of-channel emission measures the amount of interference the user cause to different users of adjacent channels in other transmitter or receiver bands. This type of emission is caused by the modulation type itself or by wideband phase noise from the PLL and other circuits in the signal chain. For the case of out-of-band emission, it is usually determined by how much interference the user can cause to all other users of the radio spectrum in other standards (e.g. military, aviation, police, etc). This type of emission is typically caused by wideband phase noise and spurs.

C) Phase Noise & OFDM

Recently, wideband digital radio systems are increasingly relying on multicarrier modulation schemes to achieve high data rates (i.e. Orthogonal Frequency Division Multiplexing, OFDM). Multicarrier OFMD currently used for WLAN and WiMAX is shown to be very robust against the common "multipath" problems since it suffers from frequency selective fading much less than single carrier systems. However, the performance of OFDM systems is shown to be more sensitive to the close-in phase noise performance of both transmit and receiver oscillators compared to single carrier systems. In a single carrier narrowband system, phase noise occupies a small part of the total bandwidth.

While in OFDM, each sub-carrier contributes its own phase noise to the overall modulated waveform. The thermal and 1/f noise in the LO circuit elements generate sideband phase noise that affects the OFDM signal during the modulation and demodulation process and this reduces the BER of the signal6, as shown in Figure 8. Carrier phase noise impacts the system BER in two ways: 1) common phase error (CPE) and 2) inter-carrier interference (ICI). CPE is caused by self-mixing of each sub-carrier with the low frequency part of its own phase noise spectrum. The effect of CPE is that all sub-carriers are rotated by a common random phase angle. Since all carriers suffer the same CPE, it is possible to both measure it and correct it in baseband using special pilot sub-carriers. The ICI is caused by the mixing of the phase noise sidebands of all neighboring sub-carriers with the desired sub-carrier. Both close-in and far-out phase noise of the LO contribute to ICI. The SNR loss due to ICI is a function of the oscillator phase noise profile and sub-carrier spacing. For a fixed channel bandwidth, decreasing the sub-carrier spacing (or alternatively increasing the number of sub-carriers) will result in rapid loss of SNR. In a typical OFDM system, the PLL phase noise is integrated over a specified band and needs to be below the EVM requirement of the system. For example, in the WiMAX standard, the integrated phase noise is < 1 degree (rms) with an integration frequency of 1/20 of the tone spacing to 1/2 the channel bandwidth. Therefore, the phase noise integration period can start as low as $100Hz$ for the smallest channel bandwidth case of $1.25MHz$. In order to meet such a requirement, all the PLL sub-circuits (i.e. charge pump, VCO, LPF, etc) need to have very low 1/f phase noise contribution. Table 3 can further be used to work out the total integrated phase noise in dBc that meets a given rms phase error requirement.

3.2   Lock Time

The lock time is the time that it takes the PLL to switch between two different frequencies. This time is measured from the start of the frequency switching action to the time the new frequency settles within a specified accuracy. When the PLL switches between two different frequencies, the chip can neither transmit nor receive any data until the frequency offset error is acceptable. In turn, this reduces the effective data rate that the system can achieve. While the PLL lock time is mainly dependent on the loop bandwidth, it also depends on the size of the frequency jump during the PLL switching. A rough number that can be used to estimate the PLL lock time is:

$$LockTime \approx \frac{3}{loopBW} \qquad (12.13)$$

Phase noise sidebands from adjacent sub-carriers contributes noise ICI to the desired carrier

Oscillator phase noise of the sub-carrier effecting itself (known as CPE)
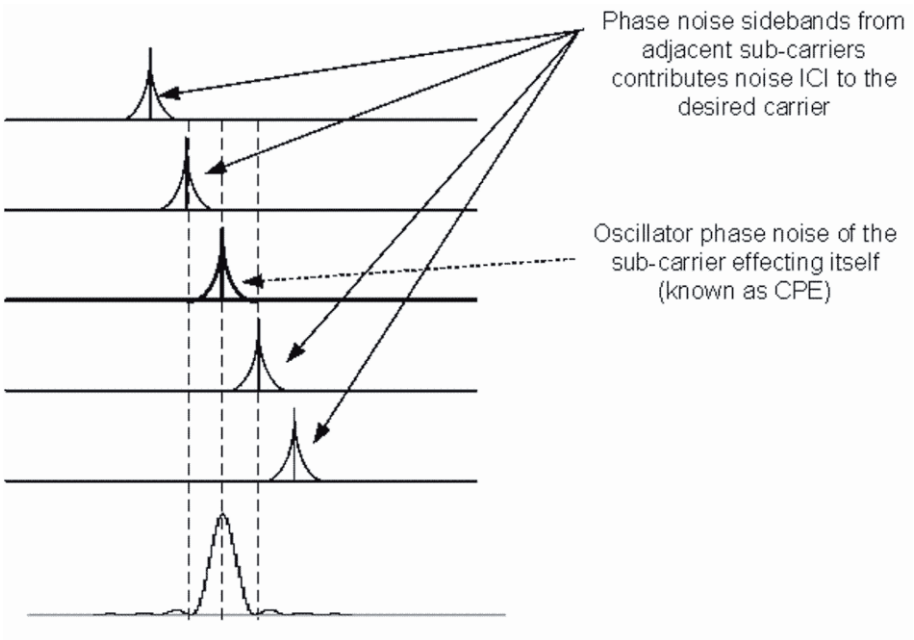
Figure 12.8. The effect of oscillator phase noise sidebands on the modulated sub-carrier in OFDM systems.

Table 12.3. Relation between integrated phase noise and rms phase error. RMS Phase Error.

RMS Phase Error (Φ_{rms} indegrees)	Integrated Phase Noise (dBc)
5	-21.2
4	-23.1
3	-25.6
2	-29.1
1.5	-31.6
1	-35.1
0.8	-37.1
0.6	-39.6
0.4	-43.1

In general, the PLL lock time is driven by the given standard specifications. For example, the Bluetooth standard employs a transmission technique called frequency hopping spread spectrum whereby the carrier frequency of the transmitter changes channels up to 1600 times per second. The PLL then needs

to switch frequencies every $625\mu s$, which means its lock time cannot exceed a fraction of this time. In WLAN, during receive/transmit switching, the perturbation to the VCO can temporally drive the PLL out of lock. This in turn can generate a transient LO frequency offset that affects the accuracy of the frequency estimation loop in the baseband MODEM, thus degrading the total system performance. To satisfy the TX to RX switching time requirement of the WLAN standard while budgeting for this frequency offset, the PLL lock time is specified to be $< 10\mu s$. In the GSM standard, while the lock time requirement is also dictated by the TX to RX switching time, it is relaxed to $< 200\mu s$. In multi-band radio, designing a single PLL that satisfies a wide variation in lock times is a formidable task that may even be impossible without programming the PLL loop filter. For example, an attempt to force the tight lock time requirement of WLAN on all standards translates into wider PLL loop bandwidth (BW), e.g. in the order of $1MHz$. However, this is far from optimum when it comes to meeting the much tighter in-band phase noise specification for GSM. On the other hand, satisfying the GSM phase noise mask typically requires a loop BW in the $100KHz$ range. This will then not meet the WLAN switching time requirement. To address these conflicting requirements, having a programmable loop filter for each standard can be an easy solution, but it will definitely require large Si die area to accommodate for the different capacitor sizes in each LPF setting. Another alternative is to use a digital LPF, which can be easily programmable to address different standards. This requires a different PLL architecture that will be further discussed in the following sections.

3.3   Bandwidth

Among all the different PLL design parameters, the loop bandwidth is the most critical with respect to the PLL performance across all other parameters. The loop BW optimization process presents a tradeoff between lock time and phase noise/spurs. While having a narrow loop BW reduces the impact of phase noise and spurs, it degrades the PLL lock time. Conversely, making it wider will improve the lock time at the expense of higher phase noise and lower spur rejection. In a simple form, a PLL is treated as a continuous time system that can be analyzed in the Laplace domain. However, due to the sample nature of the phase detector the loop needs to be analyzed in the sampled z-domain. If the loop BW is set to be less than $1/10^{th}$ of the input reference frequency then it is possible to approximate the PLL by a linear function and analyze its loop response in the continuous time (i.e. s-domain). When designing the PLL's LPF, a choice needs to be made on the type and order of this filter. Figure 9 shows the topology of both active and passive type filters assuming a charge pump based PLL design. Note that the PLL is always one order higher than the LPF because the VCO performs an integration of the control voltage and thus provides a factor of $1/s$ in the loop transfer function.
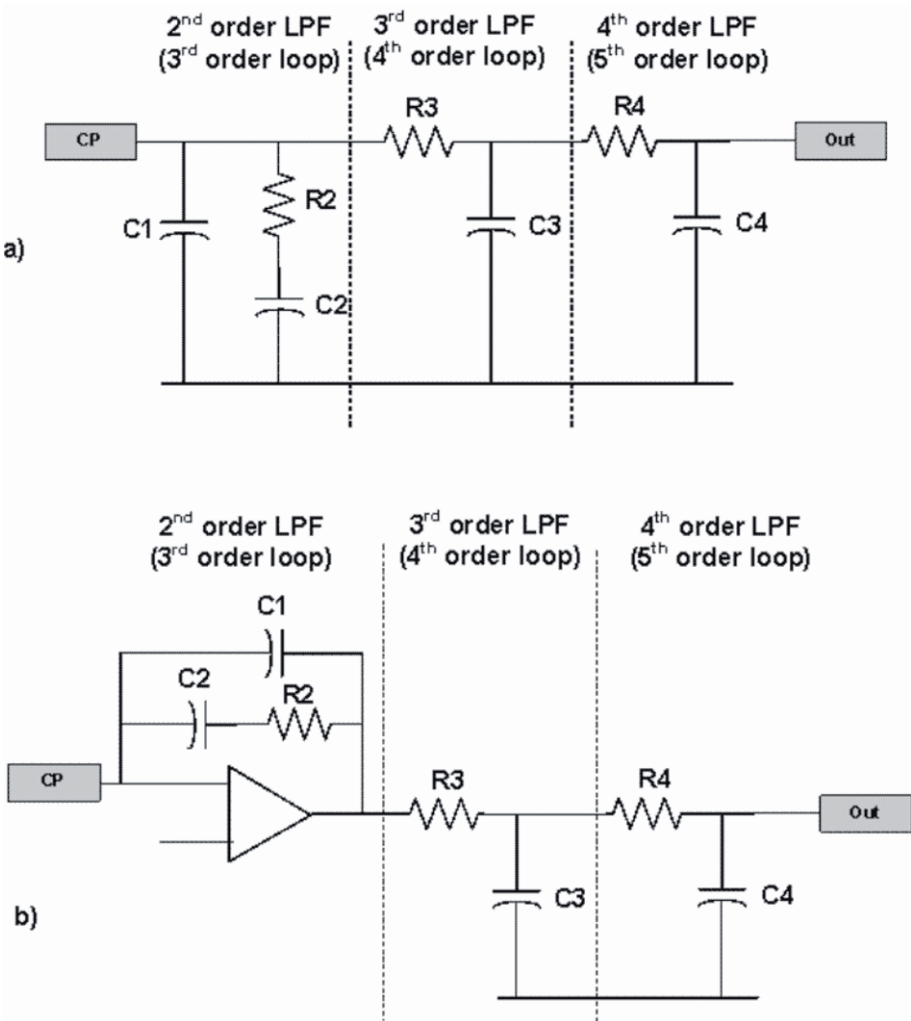
Figure 12.9. LPF types: a) Passive LPF, b) Active LPF.

The vast majority of PLLs available in the market today use a second order LPF as it provides sufficient attenuation for the reference spurs. In the wireless domain, a third or even fourth order LPF is often required to provide a superior spur attenuation level in comparison to the second order LPF. In integer-N type PLLs, both charge pump leakage and current mismatch contribute to spurs at the reference frequency and its harmonics away from the carrier signal. Fractional-

Table 12.4.   Relative spur attenuation level (dB) for 3rd and 4th order LPF relative to 2nd order LPF.

		Spuroffset frequency/Loop unity gainfrequency				
		10	20	50	100	1000
Loop Order	3	1.4	6	13.5	19.5	39.4
	4	0.6	8.2	22.5	34.3	74.2

N type PLLs suffer from the same problem in addition to fractional spurs. These spurs are generated by the tonal content of the $\Sigma\Delta$ modulator in the PLL feedback path. Another source of fractional spurs is out-of-band noise folding as a result of charge pump and phase detector nonlinearity. Both reference and fractional spurs can degrade the rms phase noise and/or the spectrum mask specifications of the PLL. The addition of one or two poles to the second order LPF can significantly reduce the level of these spurs while having a minimal impact on the loop stability, especially if the location of this additional pole(s) is chosen correctly. Typically the additional pole(s) is positioned at 10x or higher away from the loop unity gain frequency to have minimal impact on the phase margin and stability. The spur attenuation level is directly related to the position of the spur away from the loop's unity gain frequency. Table 4 lists the spur attenuation level for both 3rd and 4th order LPF relative to the 2nd order attenuation level at different spur frequencies. As the table indicates, there is little gain in spur attenuation using a 3rd or 4th order LPF compared to a 2nd order LPF if the spur frequency is below 50x the loop unity gain frequency. If the spur frequency is equal or above 100x unity gain frequency, a 4th order LPF can provide sizable advantage compared with a 3rd order LPF. Oftentimes, moving to a 4th order LPF design is the only way to get rid of a particular spur that prevents the system from complying with the spectrum mask requirement.

In terms of phase noise, the PLL loop BW can either suppress or increase the noise impact of different PLL subcomponents on the overall phase noise. Inside the loop BW, the VCO phase noise is attenuated as it gets high-pass filtered by the loop response. Outside the loop BW, the noise from the reference oscillator, charge pump, divider and phase detector is attenuated as it is low-pass filtered by the loop response. From the perspective of phase noise, deciding on the optimum loop BW requires prior knowledge of the noise from the individual PLL subcomponents. The optimum phase noise point can be reached by setting the PLL unity gain frequency to equal the interception point of high-pass and low-pass noise components. Figure 10 illustrates an example of a typical phase noise plot showing the contribution of different PLL sub-blocks and the overall PLL phase noise curve. The integrated rms phase error for each block is also
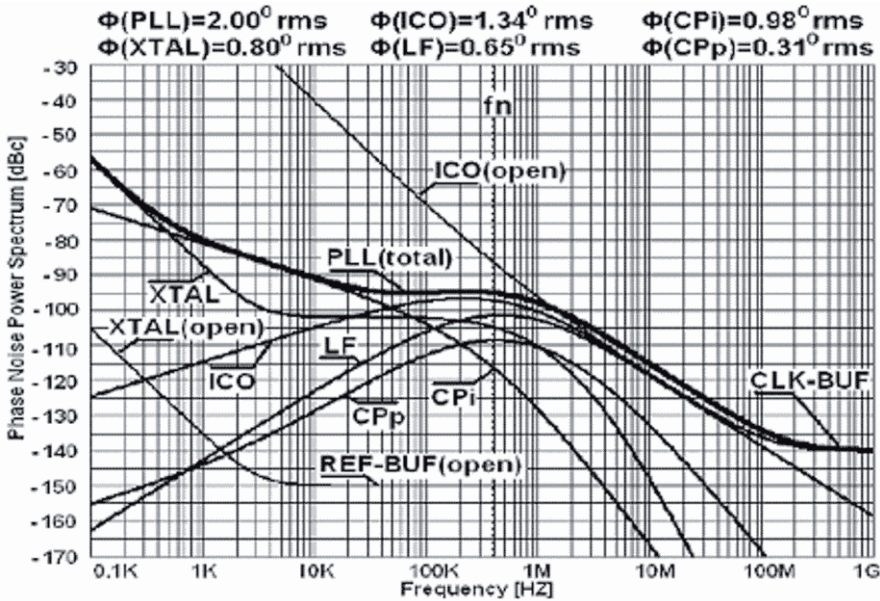
Figure 12.10. PLL phase noise example showing contribution of different PLL sub-circuits. (A. Maxim, Silicon Labs Inc.)

listed on top. The plot shows that the loop BW was optimally placed where the high-passed VCO noise (LCO) equals the dominant low-passed noise from the charge pump (CPi).

4. Impact of Technology Scaling

In this section, we will describe in detail the challenges of PLL design with technology scaling. In particular, we will focus on the degradation in PLL performance due to leakage, charge pump current mismatch and voltage headroom decrease.

4.1 Charge Pump Leakage Effects

Consider the general case of having a static phase error at the input of the PFD. Figure 11a shows the relation between the average charge pump current and the input static phase error. Assuming a certain phase error of $2\pi\tau/T$, the corresponding average charge pump current is δI. In the time domain, the charge pump output current is comprised of a train of pulses with pulse width τ, as shown in Figure 11b.

The Fourier series expression of the periodic train of pulses shown in Figure 11b can be expressed as:
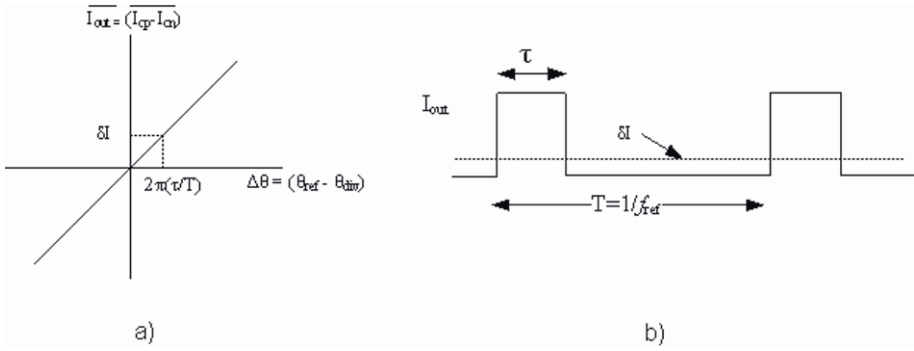
Figure 12.11. a) Average charge pump current vs. input phase error at the phase detector. b) Time domain representation of charge pump output.

$$I_{out}(t) = I_{cp}\frac{\tau}{T} + 2I_{cp}\frac{\tau}{T}\sum_{n=1}^{\infty}\frac{\sin(\pi n\frac{\tau}{T})}{\pi n\frac{\tau}{T}}\cos(2\pi n f_{ref}t)$$

$$= I_{cp}\frac{\Delta\theta}{2\pi} + 2I_{cp}\frac{\Delta\theta}{2\pi}\sum_{n=1}^{\infty}\sin c(\pi n\frac{\tau}{T})\cos(2\pi n f_{ref}t) \qquad (12.14)$$

$$= \delta I + 2\delta I\sum_{n=1}^{\infty}\sin c(\pi n\frac{\tau}{T})\cos(2\pi n f_{ref}t)$$

For small phase error ($\Delta\theta \ll 1$), Eq. (14) can be approximated as:

$$I_{out}(t) = \delta I + 2\delta I\sum_{n=1}^{\infty}\cos(2\pi n f_{ref}t) \qquad (12.15)$$

The above equation shows that the average charge pump current for the given phase error is δI, while the magnitude of the harmonics at $n f_{ref}$ is twice the average value. It is also clear that for the case of zero static phase error under lock condition, both the charge pump average and reference harmonics currents are set to zero. Now consider the case of constant charge pump leakage current (I_{leak}). Figure 12a shows the output of the charge pump current at a given offset voltage. In order for the loop to lock under zero static current condition, the turn on time for the up current is increased for a small period of time τ, as shown in Figure 12b. This is further illustrated in Figure 13 where- unlike the ideal lock situation, the charge pump is locked at a phase offset $2\pi\tau/T$ corresponding to the leakage current I_{leak}. Referring back to Eq. (15), the charge pump current as a function I_{leak} can be expressed as:
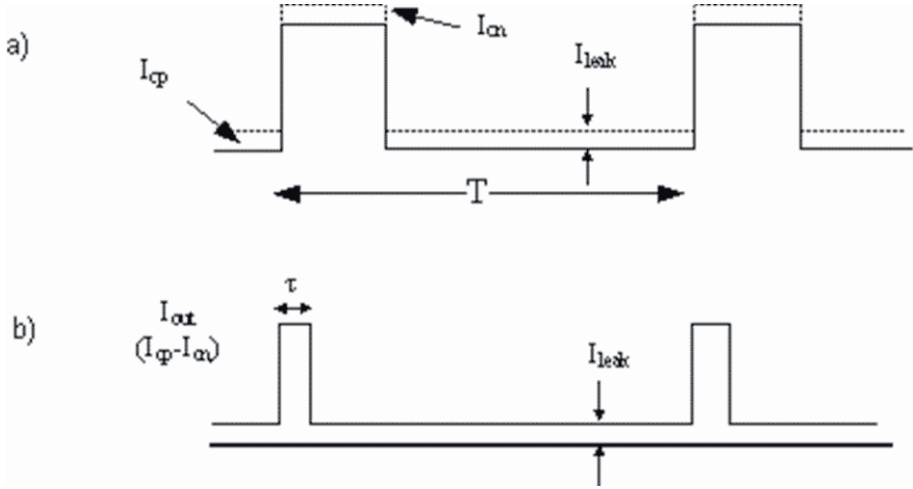
Figure 12.12. a) Charge pump up and down current at arbitrary phase offset showing the leakage effect. b) Charge pump output under lock condition.

$$I_{out}(t) = I_{leak} + 2I_{leak} \sum_{n=1}^{\infty} \cos(2\pi n f_{ref}t) \qquad (12.16)$$

The next step is to calculate the frequency deviation error $\Delta f(n)$ at the nth harmonic of the reference clock:

$$\Delta f(n) = 2I_{leak}Z_{filt}(nf_{ref})K_v \qquad (12.17)$$

The peak phase deviation can then be expressed as:

$$\Delta\theta(n) = \frac{2I_{leak}Z_{filt}(nf_{ref})K_v}{nf_{ref}} \qquad (12.18)$$

And the relative spur level to the carrier signal is related to the peak phase error by:

$$L(nf_{ref}) = 20\log\left(\frac{\Delta\theta(n)}{2}\right) = 20\log\left(\frac{I_{leak}Z_{filt}(nf_{ref})K_v}{nf_{ref}}\right) \qquad (12.19)$$

Now consider as an example a 2nd order LPF, the filter impedance Z_{filt} can be expressed as:

$$Z_{filt}(s) = \frac{1 + sR_1C_1}{s(C_1 + C_2)(1 + sR_1\frac{C_1C_2}{C_1+C_2})} \qquad (12.20)$$
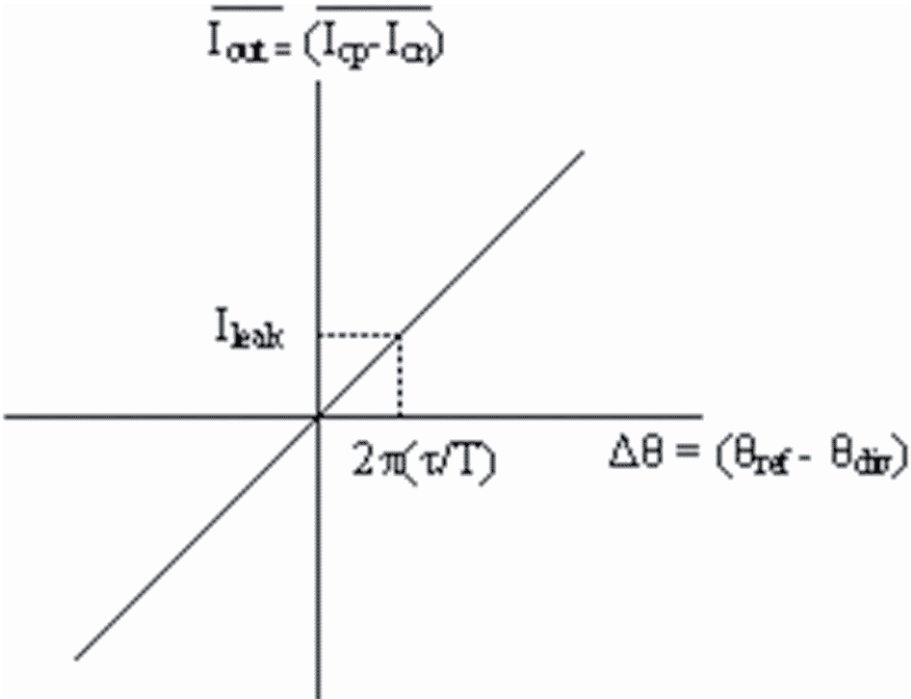
Figure 12.13. Charge pump linear phase analysis showing locking under leakage condition.

Substituting Eq. (20) back into Eq. (19) and assuming that $f_{ref} \gg unitygain$ $frequency$, the relative spur level can be approximated as:

$$L(nf_{ref} = 20\log\left(\frac{I_{leak}K_v}{2\pi nC_2f_{ref}^2}\right)$$

(12.21)

A plot showing the reference spur level as a function of the leakage current for a 2nd order PLL is shown in Figure 14. For this PLL, the reference frequency is set to be 50x the unity gain frequency. The plot shows a spur level of $-53dBc$ for leakage current set to 1% of the charge pump current. This clearly underscores the fact that even a small amount of leakage can have a large impact on the spur levels.

4.2   Charge Pump Mismatch Effects

The effect of charge pump current mismatch is shown in Figure 15a. The time ΔT is defined as the finite pulse width for the phase detector reset signal. In the majority of PLLs, this represents an intentional delay added in the reset path to remove the dead-zone problem in the PFD. The difference between the
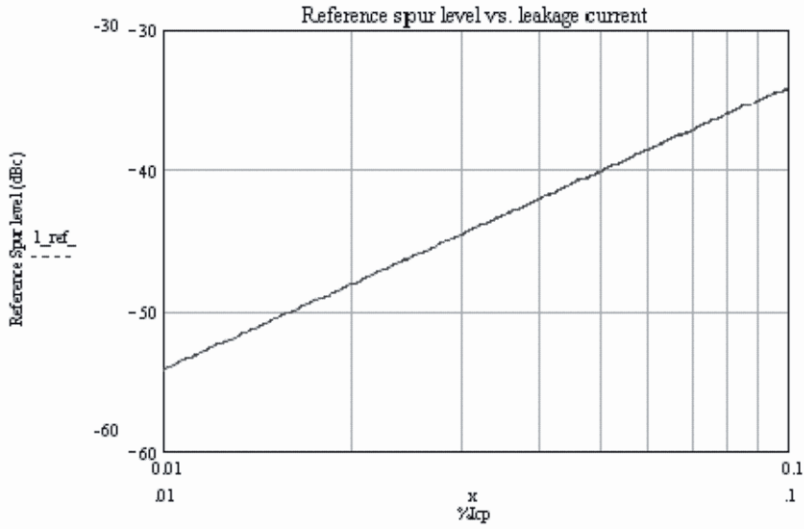
Figure 12.14. Reference spur level as a function of leakage current.

charge pump mismatch and leakage effects is that the later is active during the entire period while the former is only active during the switching time of the charge pump currents. The value of the mismatch current between down and up currents is defined as ΔI. Figure 15b shows the output of the charge pump under lock condition. The up current signal arrives slightly ahead of the down signal by time τ to cancel the effect of the mismatch current. This τ can be calculated by equating the average of the up and down pulses:

$$I_c\tau = \Delta I(\Delta T - \tau) \rightarrow \tau = \frac{\Delta I}{I_c + \Delta I}\Delta T \approx \frac{\Delta I}{I_c}\Delta T \tag{12.22}$$

The Fourier transform of the total output current can then be expressed as the superposition of two trains of pulses:

$$
\begin{aligned}
I_{out}(t) &= I_c\frac{\tau}{T} + 2I_c\frac{\tau}{T}\sum_{n=1}^{\infty}\cos(2\pi n f_{ref}t) - I_c\frac{\tau}{T} \\
&\quad -2I_c\frac{\tau}{T}\sum_{n=1}^{\infty}\cos(2\pi n f_{ref}(t-\tau)) \\
&= 2I_c\frac{\tau}{T}\sum_{n=1}^{\infty}\Big[\cos(2\pi n f_{ref}t)(1-\cos(2\pi n\frac{\tau}{T})) \\
&\quad -\sin(2\pi n f_{ref}t)\sin(2\pi n\frac{\tau}{T})\Big]
\end{aligned}
\tag{12.23}
$$
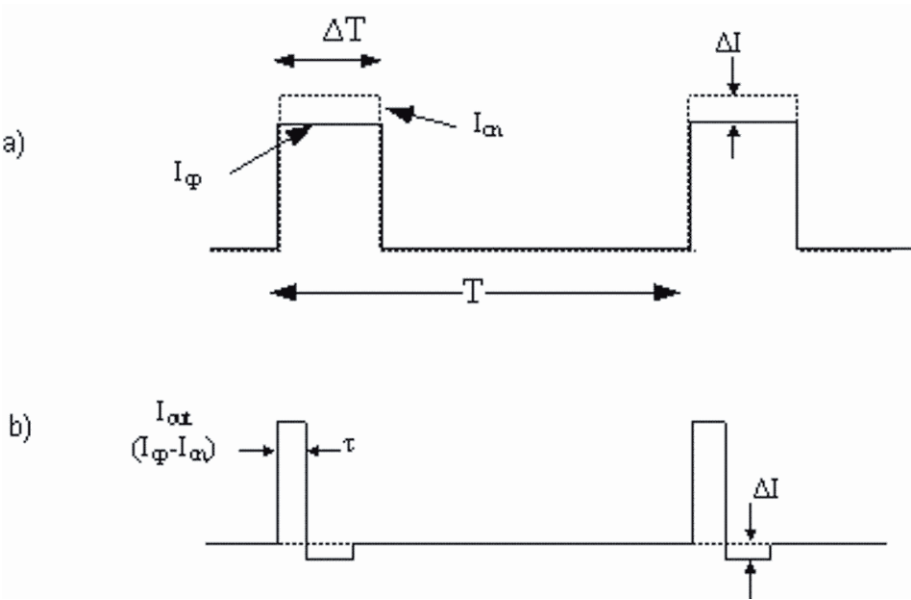
Figure 12.15. a) PLL up and down currents under mismatch condition. b) PLL output current in lock condition.

For $\tau \ll T$, the above equation can be approximated as:

$$I_{out}(t) = -2I_c\frac{\tau}{T}\sum_{n=1}^{\infty}\sin(2\pi n f_{ref}t)2\pi n\frac{\tau}{T}$$

$$= -4\pi I_c\frac{\tau^2}{T^2}\sum_{n=1}^{\infty}n\sin(2\pi n f_{ref}t) \tag{12.24}$$

Substituting τ from Eq. (22) into the above equation:

$$I_{out}(t) = -4\pi\frac{\Delta I^2\Delta T^2}{I_cT^2}\sum_{n=1}^{\infty}n\sin(2\pi n f_{ref}t) \tag{12.25}$$

Following the same procedure in Eq. (17)-(19), the relative spur level can be calculated as:

$$L(nf_{ref} = 20log\left(\frac{2\Delta I^2\Delta T^2 K_v}{nI_cC_2}\right) \tag{12.26}$$

Figure 16 illustrates different PLL lock conditions under positive mismatch, negative mismatch and zero mismatch conditions. Although having a large reset

pulse width is advantageous in eliminating the dead-zone problem, Eq. (26) shows a quadratic increase in the spur level as the pulse width increases. A plot showing the spur level vs. the % mismatch current in the charge pump is shown in Figure 17. For the given plot, the reset pulse width is assumed to be $100ps$ which is equivalent to a few gate delays in a $90nm$ technology. Even for charge pump current mismatch that is up to 10%, the calculated spur level is shown to be extremely low (i.e. $100dBc$). Previously, it has always been assumed that the mismatch in charge pump currents is a major source of reference spurs when compared to the charge pump leakage effect. However, the above analysis shows the opposite trend as technology scales down. This is due to the fact that the mismatch related spur is highly dependant on the reset pulse width which scales down significantly as technology advances. For both present and future wireless systems, the reference clock is kept at a fixed frequency as it is tied to the current widely available crystal sources. As technology scales down, the reset pulse width also shrinks due to smaller gate delays, which reduces the ratio between the reset pulse width to the reference clock period. Since the charge pump current mismatch is only active during the reset pulse period its overall impact is reduced as technology advances. On the other hand, the leakage related spur scales up with technology as leakage current increases and therefore is a much bigger concern to the design of the PLL. To mitigate against this problem, a 3rd or even 4th order LPF is required to provide additional spur attenuation, as indicated earlier.

4.3   Voltage Headroom

Due to a lowered breakdown voltage associated with deep sub-micron processes, several DC and AC parameters inside a PLL need to be limited. In low-voltage oscillators, in order to achieve the required VCO tuning range with a limited voltage swing the varactor is usually biased where the C(V) characteristic is very steep. Due to reduced headroom and increased output swing, another limitation comes from the cross-coupled switching devices in a VCO core. In deep-submicron CMOS implementations, it is common to use a single NMOS switching pair. This minimizes the parasitic capacitances and increases the tuning range. When a p-n junction varactor is used, the device is often biased close to $0V$. At this bias point, the dependency of the varactor capacitance on its terminal voltage becomes very high. As in any non-linear system, voltage dependent parameter changes cause amplitude-to-phase (AM/PM) noise conversion. In the case of a VCO, this nonlinear effect increases the impact of several AM noise sources, including power supply, thermal, flicker and substrate noise on the phase noise floor of the VCO. Especially in lightly doped substrates associated with deep submicron designs, cross-coupling effects become even more severe. Therefore, it is critical to keep the AM to PM noise gain factor, denoted here as $K_{AM/PM}$ to a minimum. $K_{AM/PM}$ can be represented
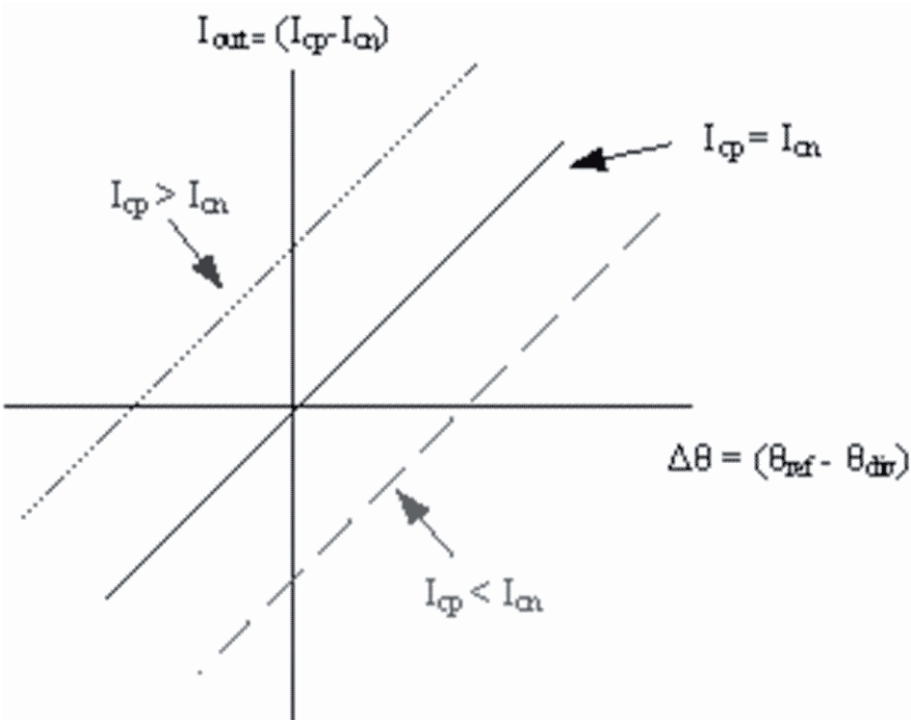
Figure 12.16.   PLL linear phase plot under different current mismatch conditions.

as:

$$\Delta\phi = K_{AM/PM} \cdot \Delta A \qquad\qquad (12.27)$$

Where $\Delta\phi$ is the phase deviation induced by amplitude deviation ΔA. This effect can be minimized by two techniques that are critical to minimize the headroom requirements in deep-submicron designs: 1) digital tuning range calibration and 2) amplitude range control.

Digital Tuning Range Calibration

Due to the increased digital gate density associated with deep-submicron processes, digital and mixed-mode calibration techniques can be utilized for increasing the tuning range of the VCO. Figure 18a shows a typical digitally calibrated VCO, where a bank of binary weighted capacitors is used for tuning the output frequency.

A segmented DAC approach can be utilized in order to cover a wide tuning range. Another advantage of the capacitor bank digital calibration approach is that it enables the use of high quality factor bondwire inductances since
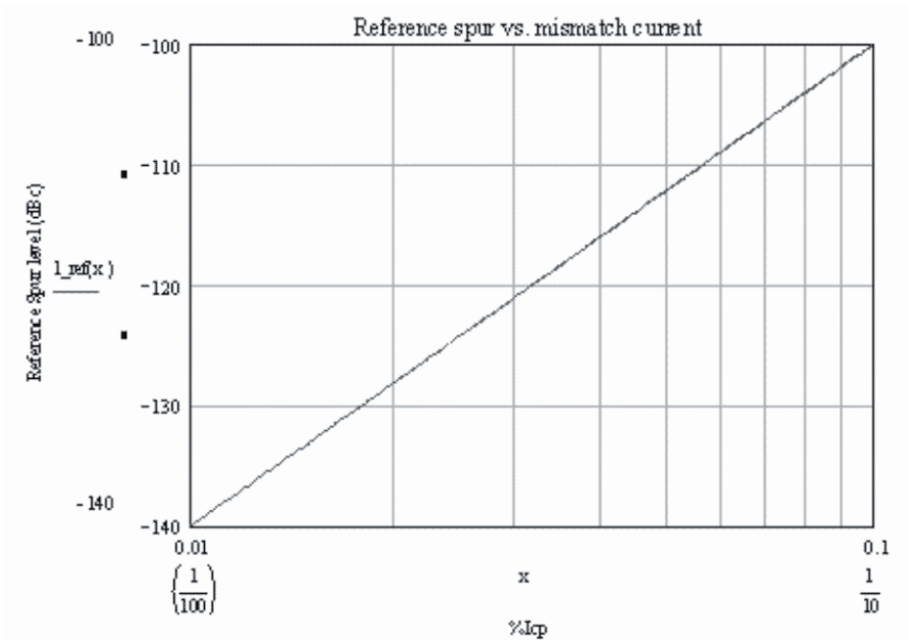
Figure 12.17. Reference spur level as a function of % mismatch in charge pump current.

manufacturing variations associated with these inductors can also be tuned by the bank of capacitors. The tuning capacitors can be divided into binary and thermometer weighted sub sections. As an example, in order to achieve 7b of capacitor tuning resolution, 4 least significant bits (LSBs) can be made up of capacitors weighted as C, 2C, 4C and 8C, whereas the remaining three most significant bits (MSBs) can be made up of seven units of 16C capacitors. A unit element of a 16C can also be divided up into two 8C units to improve matching. Typically a start-up algorithm based on a successive-approximation (SAR) technique is used to coarse and fine tune the VCO target frequency. This can be achieved by simply counting VCO edges with respect to a known reference period. In designing digitally calibrated VCOs, three critical factors should be considered: 1) effective varactor Q when the capacitor is enabled, 2) tuning ratio of the tunable capacitors, and 3) sensitivity of the tank circuit to supply and ground AM noise [5]. The coarse tuning cell is shown in Figure 18b. When B is high, M0 is on and the coarse tuning cell is on. Transistors M3 and M4 provide DC bias to ground for the drain and source of M0 ensuring minimum on-resistance $R_{ds,on}$ for M0, maximizing the effective varactors Q. When B is low, M0, M3, and M4 are off and the coarse tuning cell is disabled. Since the drain and source of M0 are now floating due to large signal swing at the VCO
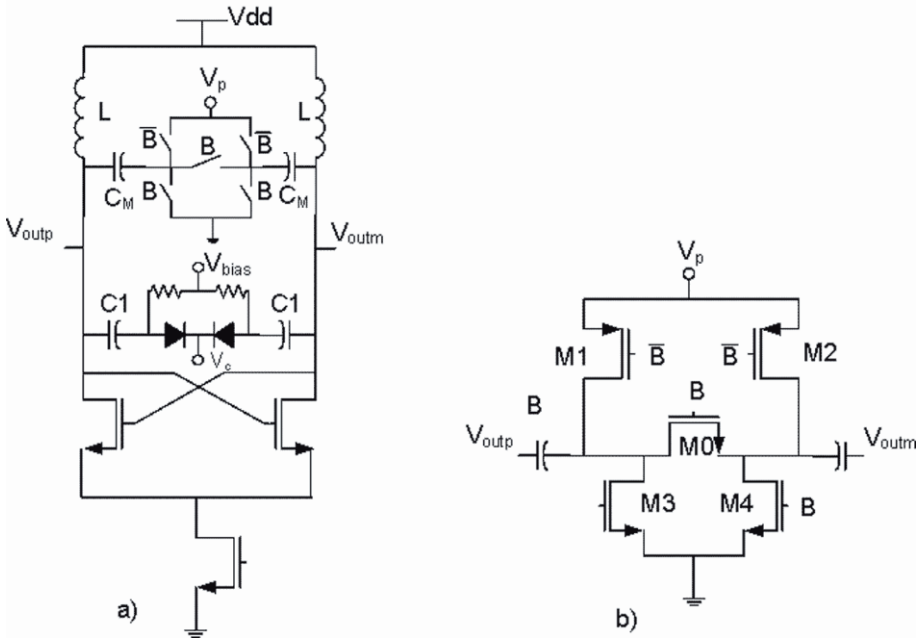
Figure 12.18. (a) A digitally calibrated VCO to mitigate reduced tuning range due to limited voltage headroom (b) Unit coarse tuning cell.

outputs, the drain and source of M0 can swing below ground and slightly turn on M0, which leads to poor off-Q. Two PMOS transistors M1 and M2 are added to bias the drain and source of M0 to V_p to ensure M0 is off. The outcome of this scheme generates a family of tuning curves for each digital setting, effectively covering an equivalent frequency range with minimum headroom requirements, as shown in Figure 19.

Oscillation Amplitude Control

In deep-submicron, wide frequency and operating range applications, the voltage swing associated with the oscillator core becomes an important reliability and power consumption concern. Another critical point to consider in deep-submicron processes is the increased flicker noise. As the oscillation amplitude increases, the flicker noise associated with the tail current source up-converts with a higher gain. For wide tuning range applications, it is critical to keep the VCO power consumption and oscillation amplitude stable [12]. This can be achieved by analog and digital means. Digital controllers have several advantages over analog ones, especially with respect to phase noise. Figure 20 shows a block diagram of a typical analog controlled VCO. The plot shows that
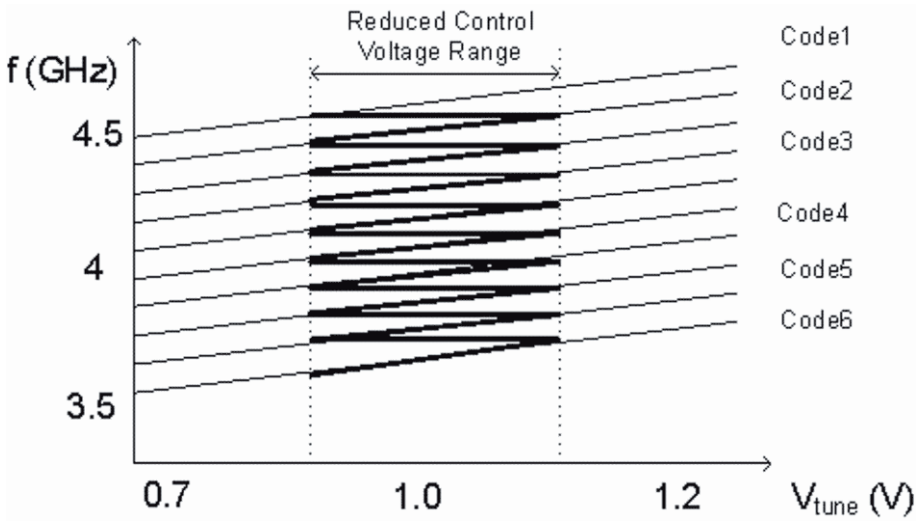
Figure 12.19. Reduced effective tuning voltage range of a digitally calibrated VCO.
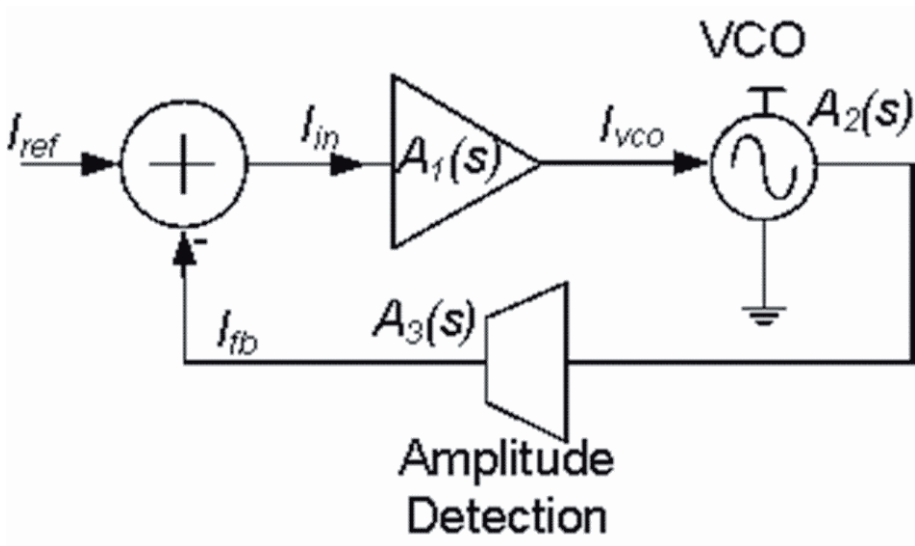
Figure 12.20. Block diagram of an analog amplitude controlled VCO.

the oscillation amplitude is continually sensed by a feedback amplifier $A_3(s)$ and error signal is fed back into the oscillator tank through $A_1(s)$.
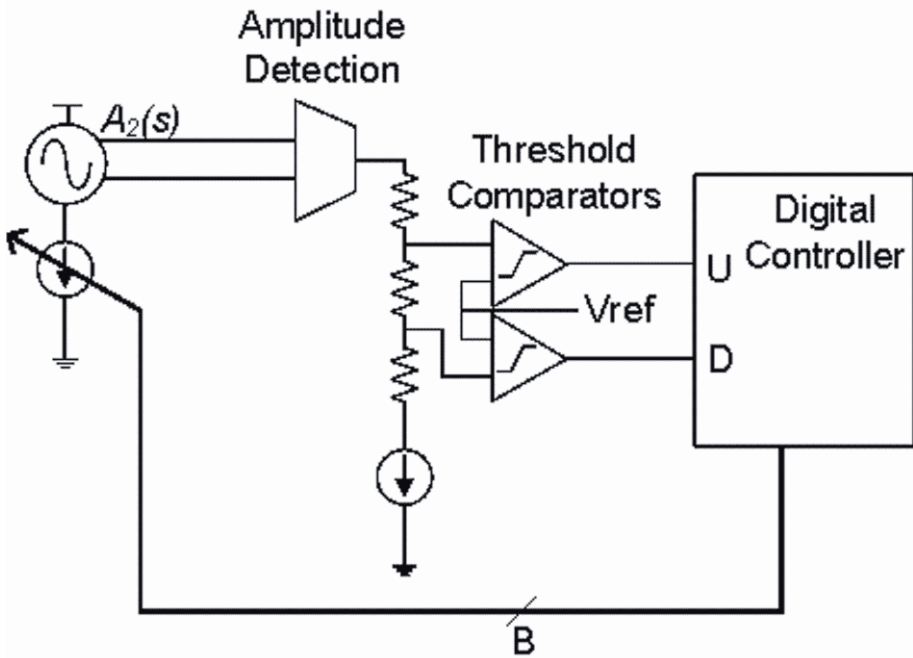
Figure 12.21.   Block diagram of a digital amplitude controlled VCO.

Due to the analog and on-line nature of this control loop, the noise associated with the feedback and error amplifiers contribute to the output phase noise. In digital applications an alternative technique can be used to continuously tune the output oscillation amplitude. Figure 21 shows a digital version of this calibration loop, where the control update is only made when needed, reducing the impact of phase noise related to the active circuitry in the control loop.

Figure 22 shows a typical application of both approaches on an LC-tank VCO. In both cases an amplitude control circuitry generates a current bias proportional to the oscillation amplitude and feeds an error signal to the oscillator core by changing its tail current bias.

5.   Architecture Landscape

5.1   Direct Digital Synthesis

Direct Digital Frequency Synthesis (DDFS) is a technology that has been in existence for nearly 30 years. Application of this technology, however, had been limited until recent years. High digital gate density in deep-submicron processes have alleviated many of the road blocks that have prevented its practical use (cost, high power dissipation, difficult implementation, and the need for
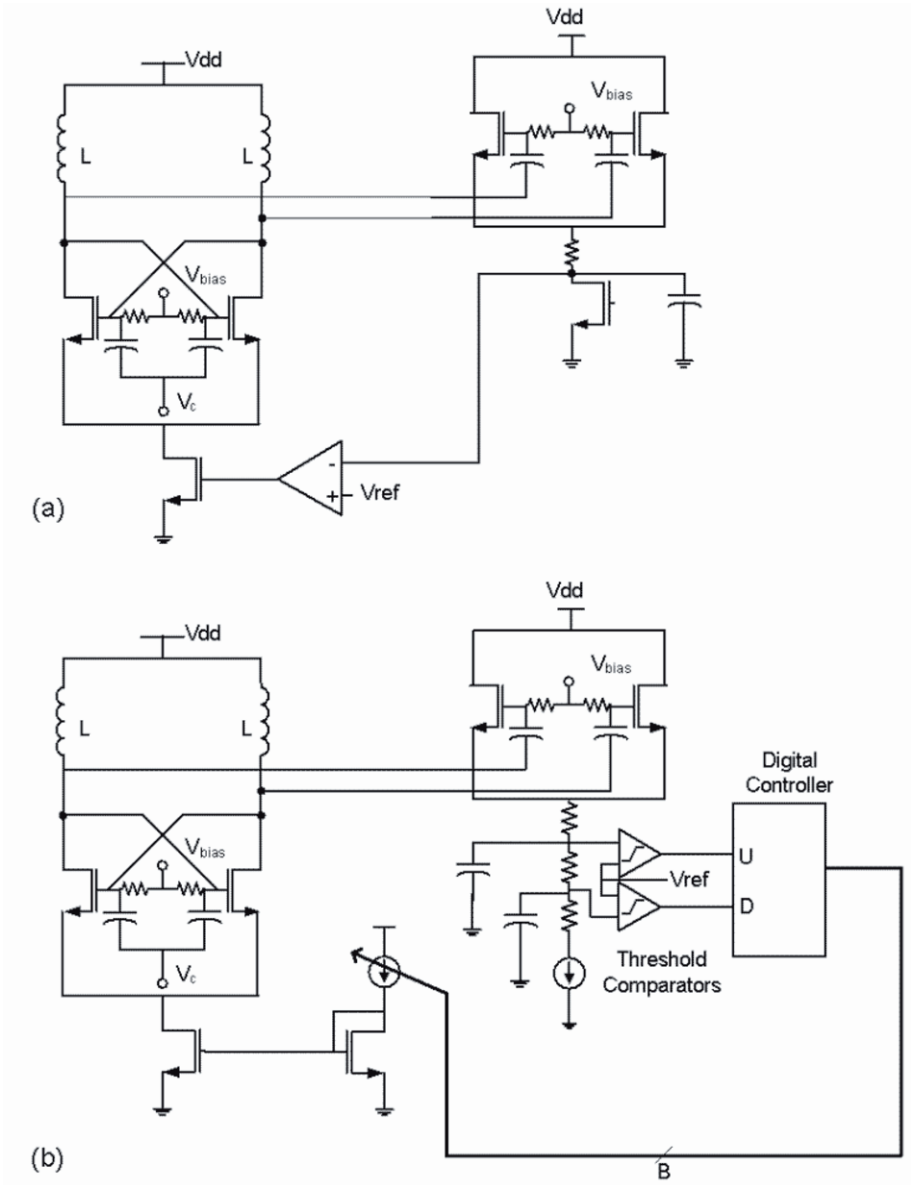
Figure 12.22. Circuit level implementation of (a) analog and (b) digital amplitude controlled deep-submicron VCOs.
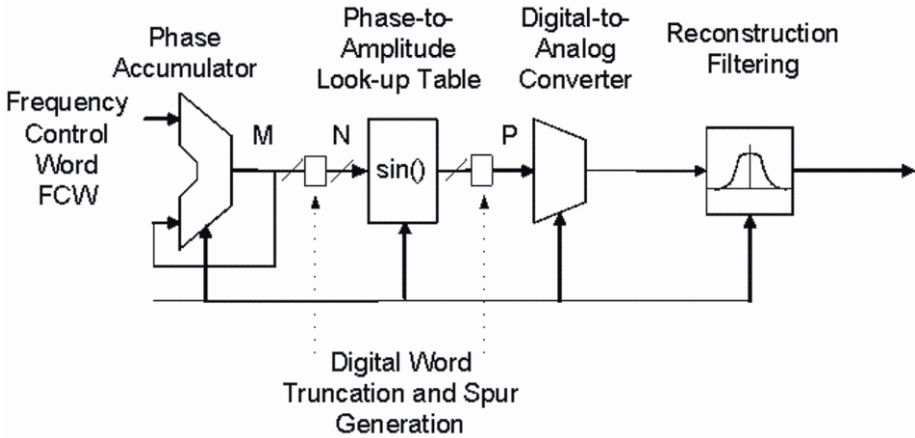
Figure 12.23. Typical DDFS with associated word truncation points and spur sources.

high speed high precision D/A converters) allowing this technology to become a very appealing alternative to analog based frequency synthesis techniques. DDFS offers some significant advantages over analog based frequency synthesis in that its programmability allows for adaptive channel bandwidths, multiple modulation formats, frequency hopping, and high data rates. These attributes have made DDFS an essential part in advancing communication system technologies. However, one drawback of DDFS for high frequency applications is that it produces high level spurious frequencies due to several numerical quantization and roundoff mechanisms inside the DDFS [14]. The conventional DDFS (sometimes referred to as the numerically controlled oscillator) consists of four basic blocks: 1) a phase accumulator 2) a phase to amplitude converter (usually a $\sin()$ look up table stored in ROM) 3) a digital to analog converter, and 4) a reconstruction filter. Figure 23 shows the block diagram of a conventional direct digital synthesizer along with digital truncation and word-limiting points that can generate spurs at the output.

The phase accumulator generates an M-bit accumulated phase information as shown below:

$$\Phi[n+1] = (\Phi[n] + FCW) mod 2^M \tag{12.28}$$

Assuming N is less than M, before going into the phase-to-amplitude conversion ROM the truncated N-bit phase would have a recursive truncation error that can be represented as:

$$ECW = FCW - \left\lfloor \frac{FCW}{2^{M-N}} \right\rfloor 2^{M-N} \tag{12.29}$$

The phase error associated with the truncated word can be represented as:

$$\Phi_e[n+1] = (\Phi_e[n] + ECW)mod2^{M-N} \qquad (12.30)$$

With the phase error shown above, the output of an ideal sin/cos look-up table becomes:

$$v[n] = \cos\left(\frac{2\pi(\Phi[n] - \Phi_e[n])}{2^M}\right)$$

$$(12.31)$$

$$= \cos\left(\frac{2\pi\Phi[n]}{2^M}\right)\cos\left(\frac{2\pi\Phi_e[n]}{2^M}\right) + \sin\left(\frac{2\pi\Phi[n]}{2^M}\right)\sin\left(\frac{2\pi\Phi_e[n]}{2^M}\right)$$

This expression generates two sideband spurs around the fundamental, with their frequency proportional to the error term $\Phi_e[n]$. As seen in this equation, this quantization error is around the fundamental and it is bandpass in nature. Another quantization error incurs at the output of the look-up table due to quantization before the DAC. This is similar to DAC quantization noise and it has a white power spectral density. The final input into the DAC can be represented as:

$$v_Q[n] = \cos\left(\frac{2\pi(\Phi[n] - \Phi_e[n])}{2^M}\right) + A_Q \qquad (12.32)$$

Where A_Q represents white quantization noise due to amplitude truncation. In order to avoid truncation problems at the $\sin/\cos$ ROM and reduce the dynamic range requirements before the DAC, noise shaping techniques similar to section 5.2 can be utilized. Figure 24 shows an application of lowpass and bandpass noise shapers to shape truncation noise into out of band quantization noise, improving overall spurious-free dynamic range (SFDR) of the DDFS.

5.2   ΣΔ Fractional-N PLLs

In conventional integer-N PLLs, the reference frequency f_{ref} dictates the minimum channel spacing. This is because the output frequency can only be an integer multiple of f_{ref}. For PLLs that require small channel resolution only small reference frequencies can be used. This results in a very narrow loop BW as the PLL BW usually tracks f_{ref} for both stability and minimum reference spurs attenuation requirements. Reducing the loop BW is undesirable as it leads to longer settling time and large area capacitors. Another important factor to consider is that all of the PLL's in-band phase noise gets multiplied by $20\log(N)$, where N is the division ratio. This could mean significantly higher phase noise when the output frequency is high and channel spacing is small. For example, in the case of GSM, the divider ratio will be as high as
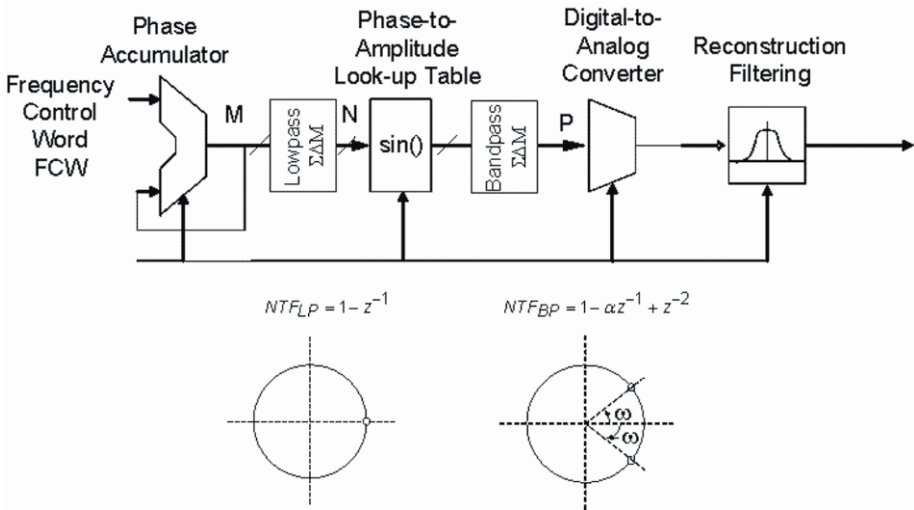
Figure 12.24. A ΣΔDDFS with a lowpass noise shaper before the amplitude LUT and a bandpass noise shaper before the DAC. Associated noise shaping functions and NTF zeros are shown below the DDFS.

10,000, which will add up to $80dB$ to the close-in noise floor. The fractional-N architecture [10] allows frequency resolution that is a fractional portion of the reference frequency, f_{ref}:

$$f_{out} = N f_{ref} \quad : \quad N = n + \frac{k}{f} \quad k = 0, 1, 2, ... \qquad (12.33)$$

where n is the desired integer part, f is the fractional resolution which is typically set as 2^m with an m-bit digital word. Therefore f_{ref} can be set to be much higher than the channel step size and overall division ration N can be reduced substantially. For the case of GSM, this means that a much higher f_{ref} can now be selected (typically $26MHz$), which implies a reduction in in-band phase noise by a factor of $42dB$ ($20 \log[26e + 6/200e + 3]$). The basic architecture of a fractional-N PLL is illustrated in Figure 25a. The architecture is very similar to integer-N with the addition of an accumulator and modulus divider (N/N+1). The main function of the accumulator is to dither between the two division values, N/N+1, to provide an averaged divide ratio that is a fractional number between N and N+1. For example, assume that the synthesizer divides by N +1 every M cycles and by N the rest of the time. The average division ratio is $N_{avg} = N + 1/M$:

$$f_{out} = \left((N+1)\frac{1}{M} + N \left(\frac{M-1}{M} \right) \right) f_{ref} = \left(N + \frac{1}{M} \right) f_{ref} \quad (12.34)$$

This is further illustrated in Figure 25b where both N and M are set equal to 4 resulting in a fractional division of 4.25. The problem with this architecture is that a periodic sawtooth waveform develops at the output of the phase detector. This periodic signal modulates the VCO creating the so called fractional sidebands (spurs). A worse case scenario can occur if a near-integer fractional setting is picked for the synthesizer. This results in fractional spurs that occur either near or inside the PLL loop BW experiencing little or no attenuation. Many methods have been developed to combat this problem such as: 1) fractional compensation by phase interpolation or 2) noise shaping $\Sigma\Delta$ modulators. Phase interpolation extracts the exact component of the instantaneous phase error signal represented by the residue signal of the accumulator [6]. This signal is converted to analog by a D/A and then subtracted from the phase detector output, which results in eliminating the spurious phase error signal. The performance of this method is highly dependent on the precise matching between the D/A converter gain and PFD gain and exact timing synchronization between the error signal path and compensation path. Because this is highly dependent on analog component matching, fractional compensation is usually not perfect. Another problem with fractional compensation is that it tends to raise the phase detector noise floor through the additional D/A noise. It is thus more advantageous to prevent the generation of spurious tones rather than generating and then compensating for them.

A $\Sigma\Delta$ modulator can be used to generate a noise shaped random signal to control the modulus divider, as shown in Figure 26. Unlike the previous accumulator method that generates a periodic divider signal, the $\Sigma\Delta$ modulator generates a control sequence with an average density equal to the desired count. However, it achieves this in such a way that all phase noise is pushed to higher frequencies12. This out-of-band quantization noise is then suppressed by PLL's LPF before it gets to the synthesizer output.

In typical $\Sigma\Delta$ A/Ds and D/As, the input to the modulator is a busy AC signal. This produces an output signal with white quantization noise shaped by the modulator noise transfer function (NTF). Typical fractional-N synthesizers are used for a channel select application, where a fixed division ratio (i.e. DC-signal) is applied to the input side. This can lead to large spurious tones at the output of the modulator and hence rigorous simulations across all allowable fractional ratios are required to check for tonal content. To overcome this problem, the modulator is realized using an architecture that minimizes DC tones. Higher order (i.e. > 2) modulators produce far less spurious tones and in-band noise than 1st and 2nd order modulators. However, the modulator's out-of-band phase noise rises at a rate of $20*(L-1)dBc/dec$, where L is the
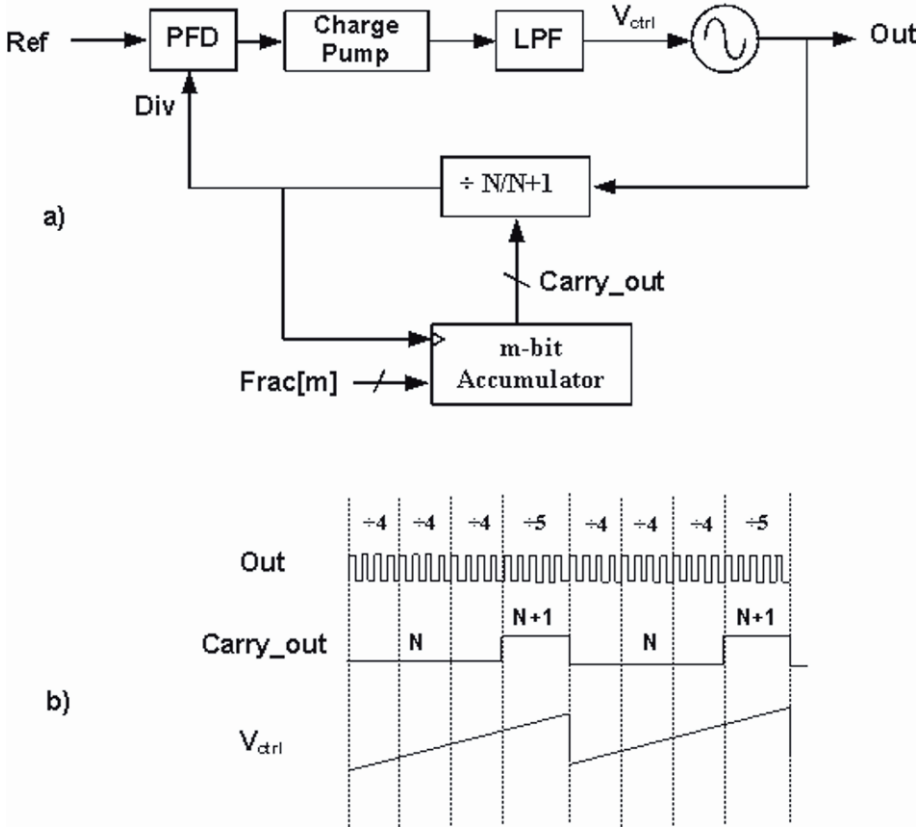
Figure 12.25. a) Basic architecture of fractional-N PLL, b) Fractional division example, N=4.25.

modulator order. Hence a LPF with order $> L$ is required to filter down out-of-band noise by at least $20dBc/dec$. Third order $\Sigma\Delta$ modulators are commonly used in today's fractional-N synthesizers as they provide low in-band noise and good tonal behavior. A single or multi-bit dithering is normally added to the input of the modulator signal to further suppress any tonal content. Figure 27 shows a block diagram of an all digital third-order $\Sigma\Delta$ modulator based on a MASH architecture. The signal $x[n]$ is typically a 16-24-bit high precision digital signal representing the desired division ratio. The output signal $y[n]$ is a 3-bit coarse quantization representation of the signal $x[n]$. The quantization error signal $e[n] = x[n] - y[n]$ is assumed to be a white noise signal that is highly shaped by the modulator's NTF. The NTF for a third-order MASH modulator can be represented by the ideal third order high-pass function, $(1 - z^{-1})^3$. The PSD of the modulator NTF is plotted in Figure 28 where the time domain noise
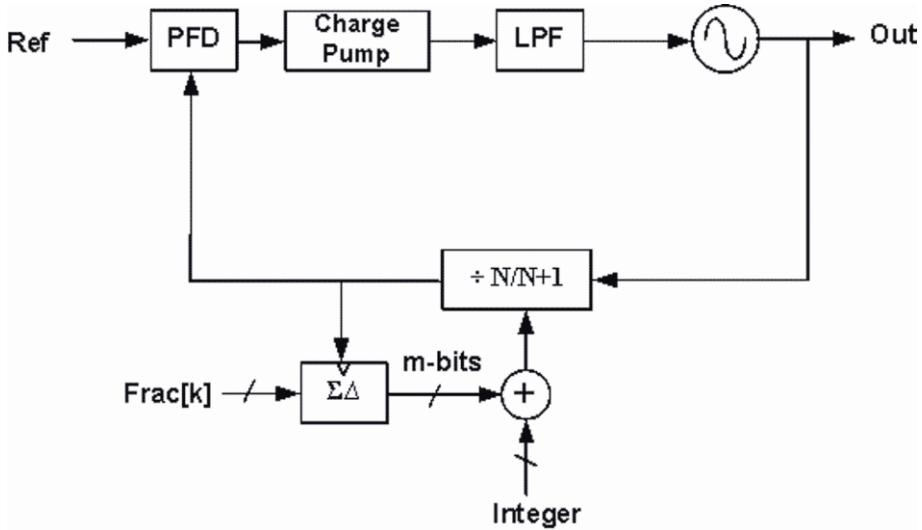
Figure 12.26. ΣΔ fractional-N synthesizer.

is shown in blue while the ideal $(1 - z^{-1})^3$ transfer function is shown as a dashed line.

Although ΣΔ fractional-N synthesizers are by far the most widely used architecture in all of today's wireless transceivers, they still have a fundamental challenge in meeting the spur level requirements in various standards. Fractional PLLs inherently suffer from two different types of spurs: 1) ΣΔ quantization spurs and 2) spurs based on out-of-band noise mixing and intermodulation. Quantization type spurs are easily predictable by simulating the tonal content behavior of the modulator. By using a simple linear model for the PLL, the simulation can show the exact amplitude and frequency of these types of spurs. The second type of spurs which are also known as near-integer fractional spurs are due to the loop high sensitivity to charge pump and PFD non-linearity [7]. These spurs simply don't show up in the PLL spectrum using just a linear simulation model. The frequency of these spurs lie at fractional frequency Δf and its harmonics from the carrier:

$$\Delta f = |f_{out} - N f_{ref}| \tag{12.35}$$

These spurs experience little or no attenuation if the spur frequency Δf lie inside or near the loop BW. This can result in the PLL failing the spectrum mask and/or EVM requirement. A simple way to reduce these spurs is by shifting the PFD operating point to avoid the charge pump non-linearity around zero phase error. This can be done by using a small but constant leakage current
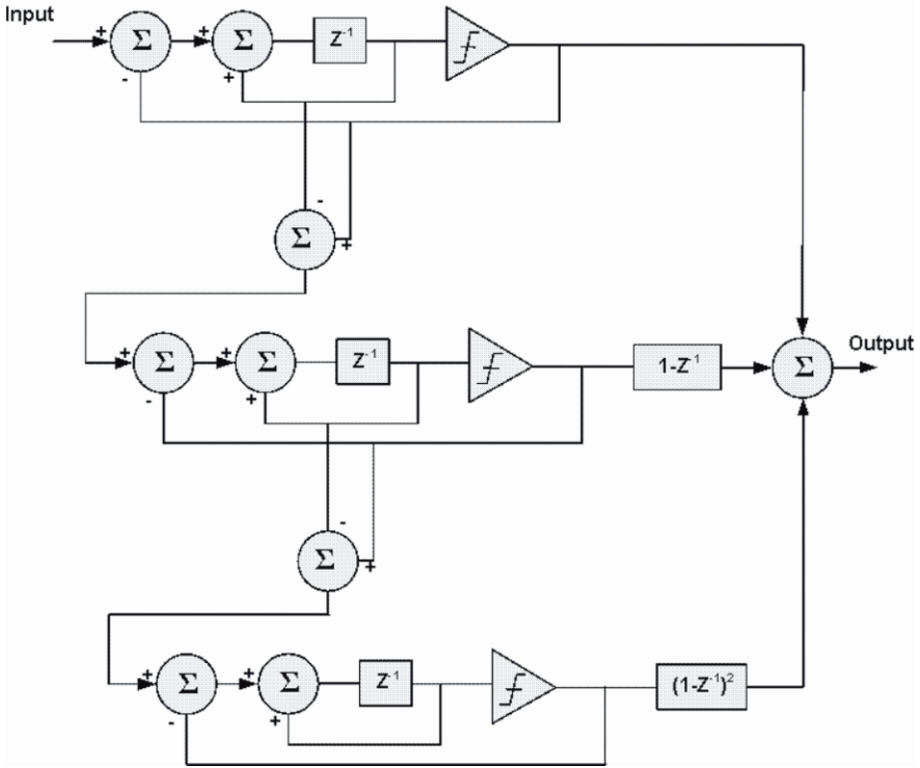
Figure 12.27.   Third-order $\Sigma\Delta$ modulator block diagram.

to force the charge pump operation outside the dead zone region. Another way to get around this problem is to use some charge pump linearization techniques that reduces both transient and DC mismatch between up and down currents.

5.3   All Digital PLL (ADPLL)

All-digital PLL (ADPLLs) have been utilized in video communications, microprocessor clock generators and in read-channel applications due to their reduced die size and fast frequency adaptation. The main difference between a digital phase detector PLL and an ADPLL is the implementation of the loop filter and in some applications, the oscillator. ADPLLs utilize a digitized version of the phase difference between the reference and feedback edges, and use minimum-phase infinite-impulse-response (IIR) digital filters for the loop filter implementation. ADPLLs can use a digital codeword controlling a digitally controlled oscillator (DCO), or a DAC to control a VCO. As discussed in the earlier sections, reduced supply voltage, wide process variations and expensive
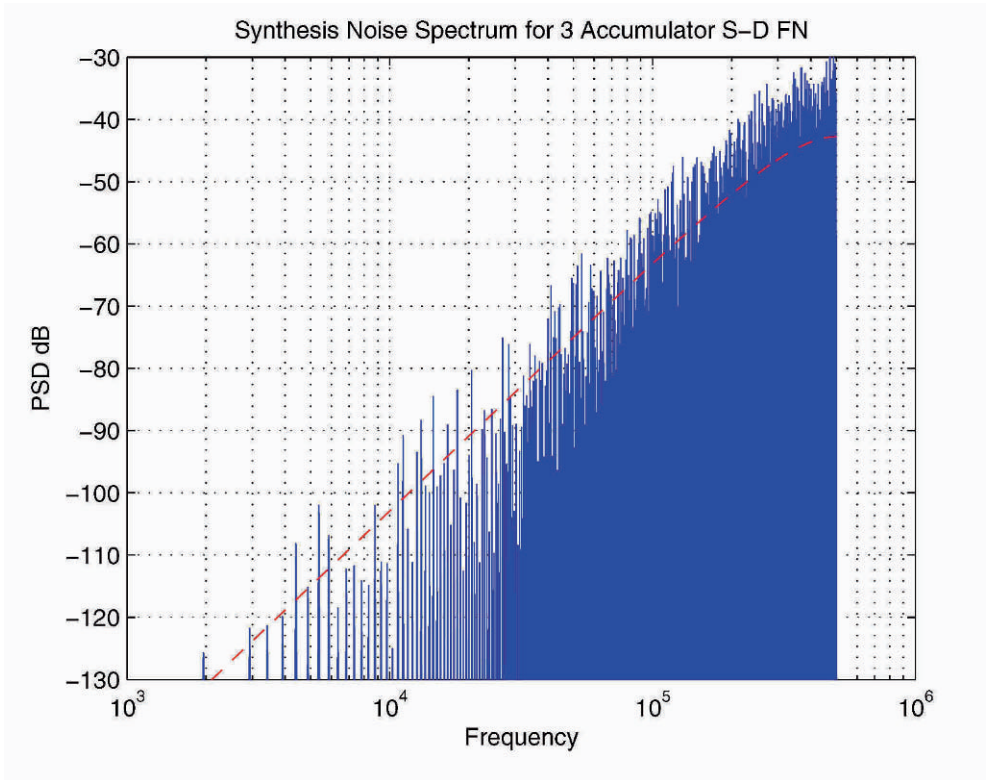
Figure 12.28. Noise Spectrum for 3rd order $\Sigma\Delta$ MASH modulator.

silicon estate associated with deep-submicron processes are all contributing factors in making ADPLLs an attractive approach for wireless and wireline communication transceivers. Digital loop filters can enable variable loop bandwidths and modulation options for the ADPLL with minimum silicon area. The ADPLLs can be used alone or with wideband analog PLLs. There are several design concerns that need to be addressed while designing ADPLLs. Due to the digitized nature of the phase detector, as in any ADC, the nonlinearity and supply rejection of this cell becomes a critical design issue. Especially in delay line based implementations, supply sensitivity of the ring-oscillator should also be minimized. The next two sections describe the operating principle of two commonly used ADPLLs.

DDFS Driven ADPLLs Unitizing Frequency Discriminator Feedback

As outlined above, due to loop parameter variations inside a PLL, having an all digital loop filter driving the voltage controlled oscillator (VCO) is a very attractive approach. The modulation data can either be introduced in the loop
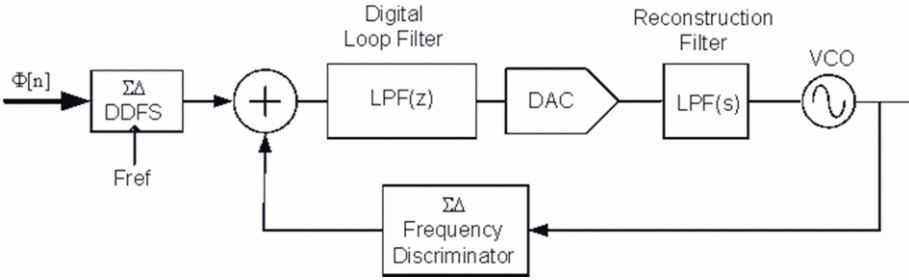
Figure 12.29.   Block diagram of a frequency discriminator controlled ADPLL.

divider, or it can be introduced at the reference frequency. A $\Sigma\Delta$ direct-digital frequency synthesizer as discussed above can be utilized to generate a modulated data as the reference frequency [13]. An alternative DDFS-like arrangement generating a first order noise-shaped single-bit output is shown in Figure 29. This $\Sigma\Delta$-DDFS can be used as a reference of an all-digital PLL, while the feedback signal comes from a $\Sigma\Delta$ frequency to digital converter ($\Sigma\Delta$FDC) [4]. Frequency to digital converters, or digital frequency discriminators are used to extract instantaneous phase, or frequency of a phase/frequency modulated signal. This building block is most commonly used at the down-converted IF frequency of a communication receiver but recently is being utilized for on-line calibration of frequency synthesizers, modulators and as a feedback divider of all-digital PLLs. The $\Sigma\Delta$FDC digitizes instantaneous frequency of a modulated waveform similar to an over-sampled ADC that processes the amplitude of a signal. For an FM modulated input signal, the input waveform can be described as

$$v_{FM}(t) = \cos\left[2\pi f_c t + 2\pi k_{FM}\int_0^t m(\tau)d\tau\right] \qquad (12.36)$$

The instantaneous frequency $f_i(t)$ is the derivative of the phase and is:

$$f_i(t) = f_c + k_{FM}m(t) \qquad (12.37)$$

where f_c is the carrier frequency, $m(t)$ is the modulating signal and k_{FM} is the frequency sensitivity (modulation index) of the modulator. A $\Sigma\Delta$FDC digitizes the deviation of $f_i(t)$ from its nominal center frequency f_c with a finite SNR and certain noise-shaping characteristics due to over-sampling. Very well-known filtering and decimation algorithms can be utilized for reconstructing the phase/frequency data out of the $\Sigma\Delta$FDM. Modulation information can be introduced into the DDFS ($\Phi[n]$) as well. As seen in Figure 29, the phase detector takes a digital difference between the input and feedback bit-streams and feeds it to the all-digital loop filter. Once lock is obtained, theoretical noise shaping response of the feedback and the reference path gets subtracted, greatly
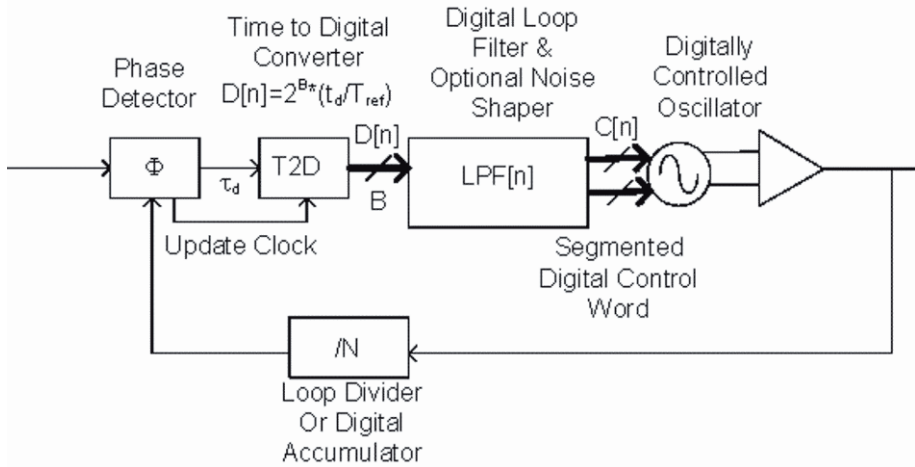
Figure 12.30. Block diagram of an ADPLL utilizing a DCO.

reducing the spurious content at the DAC output. A simple first order low-pass filter can also be used to reduce the images associated with the DAC.

ADPLLs Utilizing Digitally Controlled Oscillators (DCOs)

An alternative technique that has been utilized for ADPLLs is using a time digitizer (time-to-digital converter) to feed into a digital loop filter, eventually dithering the digital control node of a DCO, as shown in Figure 30. The DCO can be based on LC tank [15] or digital delay line based ring oscillator [8].
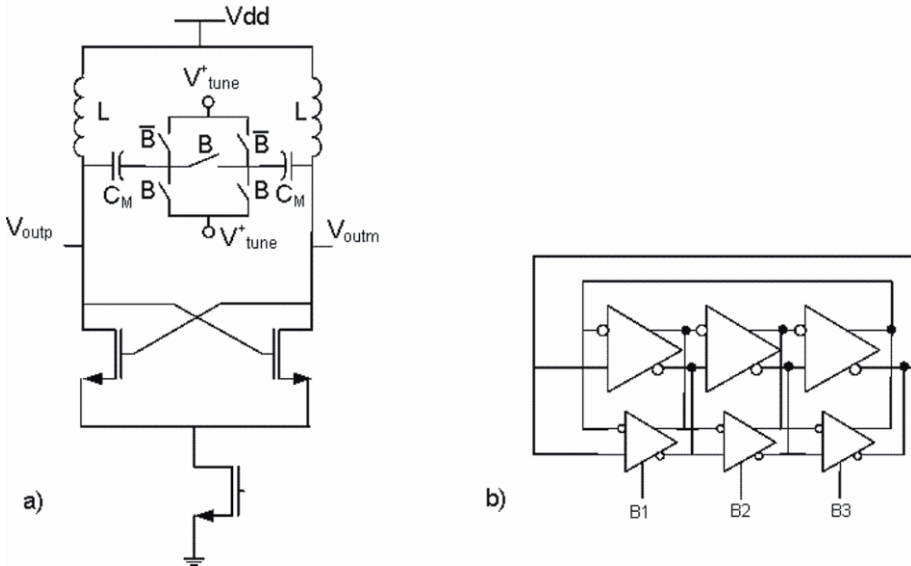
Figure 12.31. Typical DCO architectures utilizing a) programmable LC tank circuit b) programmable delay line via unit delay cells.

Two commonly used DCOs are shown in Figure 31. In the LC-tank case, the bank of unit capacitors can be dithered by the high speed control signal being generated by the digital loop filter. This effectively tunes the oscillation frequency of the LC tank. In the ring-oscillator DCO option, the unit delay elements can have adjustable drive strengths, increasing or decreasing the overall speed of the DCO.

References

[1] Bruss, S. (2001). Improving phase noise in RF Voltage Controlled Oscillators. In *http://www.uaf.edu/asgp/spbruss/other/em/phase_noise.pdf*.

[2] Crawford, J.A. (2004). Phase Noise Effects on Square-QAM Symbol Error Rate Performance. In *http://www.siliconrfsystems.com/Papers/Phase%20Noise%20Effects%20on%20Square%20QAM%20v1.pdf*.

[3] Ferre-Pikal, E. (2002). PM and AM Noise Measurement Techniques. In *Tutorial at the IEEE International Frequency Control Symposium*.

[4] Finsrud, M., Høvin, M., and Lande, T.S. (2001). Adaptive Correction of Errors in Second-Order MASH ΣΔ FDM Solution. *IEEE Trans. Circuits and Systems-II: Analog and Digital Signal Processing*, 48(11).

[5] Lee, S.T., Fang, S.J., Allstot, D.J., Bellaouar, A., Fridi, A.R., and Fontaine, P.A. (2004). A Quad-Band GSM-GPRS Transmitter With Digital Auto-Calibration. *IEEE J. Solid-State Circuits*, 39(12).

[6] Miller, B. and Conley, B. (1991). A multiple modulator fractional divider. *IEEE Trans. Instrum. Meas.*, 40:578–593.

[7] Muer, B. and Steyaert, M. (2003). On the analysis of Delta-Sigma Fractional-N Frequency Synthesizers for High Spectral Purity. *IEEE Transactions on Circuits and Systems-II*.

[8] Olsson, T. and Nilsson, P. (2004). A Digitally Controlled PLL for SoC Applications. *IEEE J. Solid-State Circuits*, 39(5).

[9] Razavi, Behzad (1998). *RF Microelectronics*. Prentice Hall PTR.

[10] Riley, T.A., Copeland, M.A., and Kwasniewski, T.A. (1993). Delta-sigma modulation in fractional-N frequency synthesis. *IEEE J. Solid-State Circuits*, 28:553–559.

[11] Robins, W.P. (1982). Phase Noise in Signal Sources. *Peter Peregrinus Ltd. (for IEE)*.

[12] Rogers, J.W.M., Rahn, D., and Plett, C. (2003). A Study of Digital and Analog Automatic-Amplitude Control Circuitry for Voltage-Controlled Oscillators. *IEEE J. Solid-State Circuits*, 38(2).

[13] Sander, W.B., Schell, S.V., and Sander, B.L. (2003). Polar Modulator for Multi-mode Cell Phones. In *IEEE Custom Integrated Circuits Conference*, pages 439–445.

[14] Song, Y. and Kim, B. (2004). A14-b Direct Digital Frequency Synthesizer With Sigma-Delta Noise Shaping. *IEEE J. Solid-State Circuits*, 39(5).

[15] Staszewski, R.B., Hung, C-M, Barton, N., Lee, M-C, and Leipold, Dirk (2005). A Digitally Controlled Oscillator in a 90 nm Digital CMOS Process for Mobile Phones. *IEEE J. Solid-State Circuits*, 40(11).

Chapter 13

RFIC DESIGN FOR FIRST-PASS SILICON SUCCESS

James Wilson and Mohammed Ismail

1. Introduction

The cost of each design cycle increases as the fabrication process technology advances. As such, new techniques are required to minimize the number of spins required before a chip meets all of its specifications. This chapter presents the motivation behind and requirements for what is termed "first pass silicon success" in the context of designing complex RF integrated transcivers for wireless applications. Design techniques leading to first pass success and taking advantage of the increased integration of digital, analog and RF are presented that address these issues.

The radio transceiver of tomorrow will be very complex, as it will have to meet increased demands of wideband operation at much higher data rates. In 4G wireless systems, convergence of cellular and WLAN traffic for VoIP will require the radio to operate in multiple RF bands and with different modulation schemes ranging from QPSK to 64- and 256-QAM OFDM. Current and future trends call for the highest levels of integration to achieve low cost and low power for handheld wireless devices. While CMOS technology scaling and innovations in platform based systems and Network-on-Chip (SOC and NOC) have resulted in great strides within the digital part (digital baseband/MAC), the radio part of a wireless solution remains a major bottleneck. In today's radio design environment, a fully integrated CMOS radio requires several silicon spins before it meets all product specifications and often with relatively low yields. This results in significant increase in NRE cost, especially that mask set costs increase exponentially as feature size scales down. Furthermore, this could lead to missing important market windows, particularly with the decreasing life cycles of semiconductor products.

Integrated RF systems lack arbitrary composability, i.e. they cannot be composed from their sub-systems as easily as their digital counterparts. RF performance is highly susceptible to random variations in process and operating conditions. Such variations do not scale with the process. Worst-case corner simulations often lead to over-design and increased power consumption. RF models, package models and design kits are based on certain assumptions that severely limit design space exploration. All these factors prohibit first-time-right silicon.

1.1   What is First Pass Silicon?

First pass silicon success means four criteria have been met:

1 Only one tapeout has been performed

2 All the blocks in the design meet their block level goals

3 The overall system specifications are met

4 The yield of the chip is at an acceptable level

Obviously for a chip to be first time right there can be only one tapeout. If it works on the second tapeout, then it is not first time silicon success! A chip cannot be considered a success if any of the blocks are not meeting their goals, or if any of the system specifications are not being met. The latter is the one that can be the toughest to meet. Chip yield is very hard to estimate for analog and RF blocks, and is usually much lower than digital yield. This will be discussed more in section 1.3.0.

1.2   Is Having Only One Silicon Spin Important?

So the question can be asked "Is having only one silicon spin important enough to worry about when there are so many other issues to worry about?" The answer is undeniably *YES!* There are three reasons that first pass silicon is important: NRE costs, mask set costs, and time to market.

NRE stands for non-recurring engineering. They are the costs associated with designing the chip such as paying the design team's salaries, design software licenses, and overhead for the project duration. Consider the typical design cycle shown in figure 13.1.

A design starts out with the system specification followed by block specifications. Once the block specifications are firm the block design start, followed by layout. There can be several loops between design and layout to account for parasitic extraction. After the layout is set, the chip is sent to fabrication. When it returns, the chip is tested to see if it meets the original specifications. If it does not meet the specs, or the specs didn't quite work well enough to pass
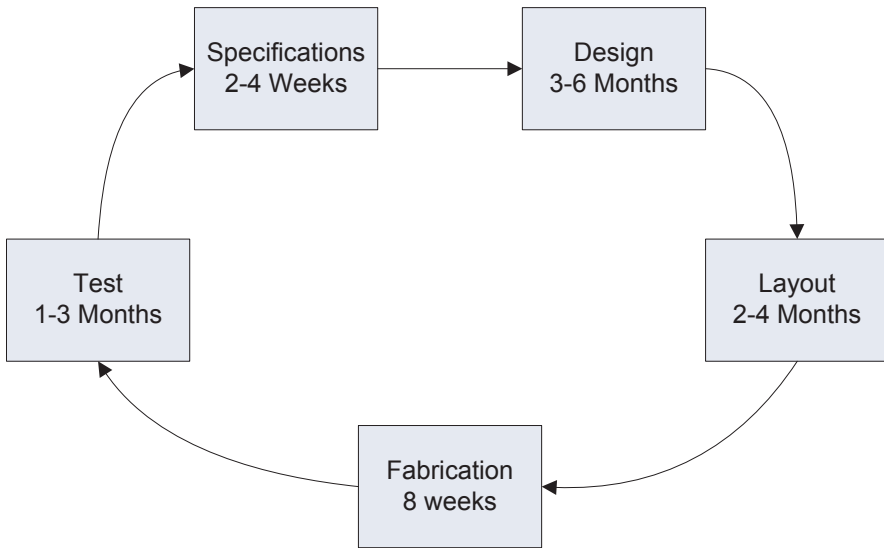
Figure 13.1. A typical design cycle can take from 6 to 9 months.

qualification, the cycle starts all over again. The first time through the design cycle can take between 6 and 9 months to complete.

The cost for a design cycle is not limited to NRE. The mask set costs are increasing exponentially for every each new process generation. Figure 13.2 plots the mask cost versus process generation. It shows that a mask set for the 90nm process costs around $900,000 currently. When moving to 65nm, the cost rises to nearly $3,000,000.

To calculate the costs of a design cycle, assume that the design team is 10 people. It takes 9 months to complete the design cycle, which is over $500,000 in NRE. At 0.18μm the cost of a mask set is $250,000. This puts the cost of a single spin at $750,000. Additional spins usually take less time and cost less, but can still cost more that $400,000. But there are other problems caused by having to respin the design.

Even if the dollar cost of extra respins can be tolerated, the company might not be able to tolerate the extra time. If the error that caused the respin is small and can be fixed in 1 month, it still takes 8 weeks to fabricate the new chip. This new chip will then have to be tested again, which is another month that the deadline slips. So in the best case, a respin costs about 4 extra months. These 4 months are time that the product is not on the market, and time that the engineers are not working on the next product.

Mask Set Costs
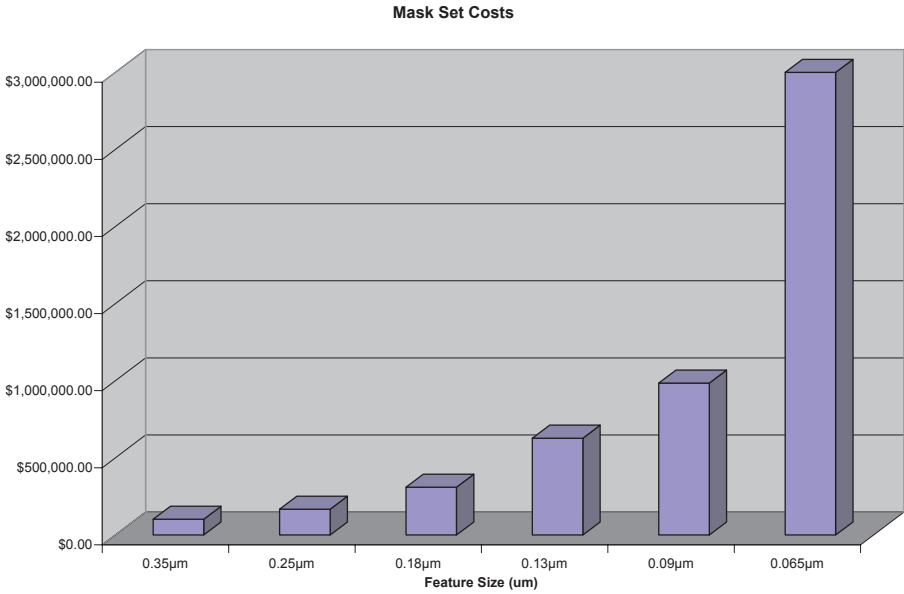
Figure 13.2.   The price of mask sets are increasing exponentially with every new process generation[1].

1.3   Why Do Chips Fail?

The causes for failure can be generalized into 5 categories:

1 Complexity

2 Errors in the system level

3 Errors in the block level

4 SoC/Integration issues

5 Yield

Many of these categories overlap with one another. Complexity is an unavoidable byproduct of integration. As system specifications become stricter, the requirements on the system level and block level become stricter.

Complexity

As an example of a complex chip, consider the chip[2] shown in Figure 13.3. This is an 802.11a/b/g WLAN radio transceiver. The chip contains 2 LNAs, 2 PLLs, 7 mixers, 2 I/Q basebands, 2 PAs, a bandgap and a crystal oscillator.

(a)

(b)
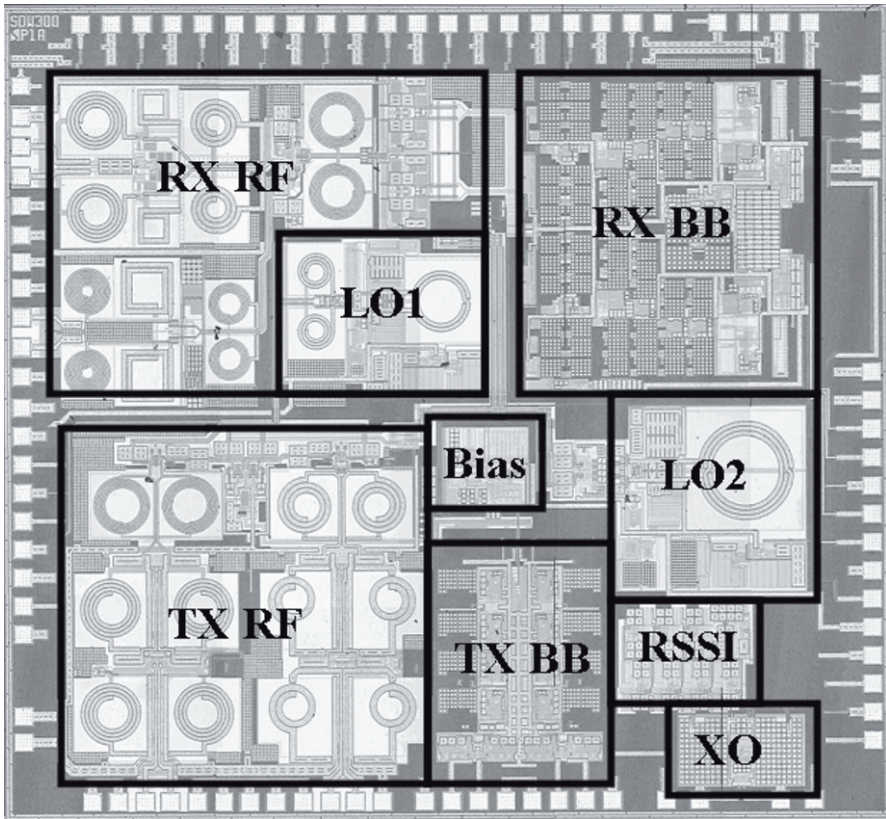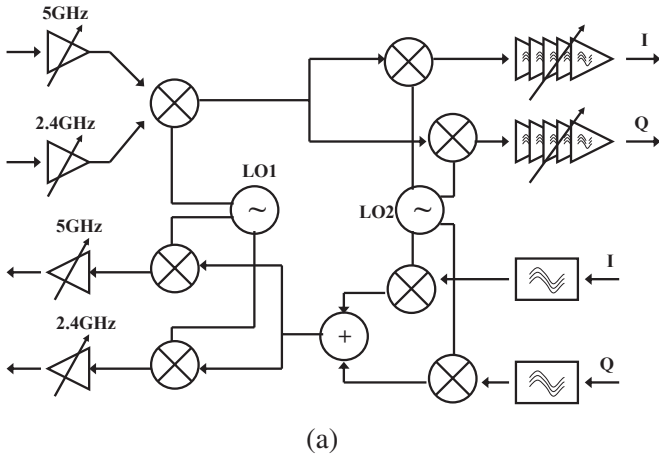
Figure 13.3. An 802.11a/b/g WLAN radio transceiver. (a) transciver architecture, (b) chip photo.

While complexity in and of itself does not cause chip failure, it can directly lead to chip failure. Assume that each block has a percentage chance to not work through some sort of error, be it in block design, block specifications, or integration issues. As chips become more complex, they include more blocks. The more blocks in a chip, the higher the odds that at least one of those blocks will not work correctly. This leads to the next cause for failure: errors at the system level.

Errors at the System Level

Once the system level specifications are stable, the block level specifications are derived. Thus if there are problems in the system level, even if the blocks meet all of their performance metrics, the chip can still end up as a failure. As an example, consider PLL phase noise. If the system level does not account for all the causes of phase noise in the complete transmit and receive chain, the PLL phase noise may end up being specified too loose. This can happen if the AM-PM noise of the PA is not taken into account, or the phase noise introduced in the digital demodulation is not handled properly. Another example of a system error is the AGC settling time. If the AGC algorithm requires 3 steps to properly set the gain, but the system specifies the group delay too high to allow 3 gain steps before data reception begins, then this is a system problem. A third example the crystal drift specifications. If the crystal oscillator drifts in a manner that is not accounted for, for example because it is a AT cut rather than a BT but, then this can lead to a much higher bit error rate than expected, even though all of the blocks are working properly.

Errors at the Block Level

There are too many ways that block design can cause failure to be able to list them all. Instead a few of the most common errors are presented here. The four most common errors in the block design are: changes in the process, parasitic elements, bad models, or incorrect simulation test benches.

While changes in the process, i.e. a process shift, cannot be seen in advance by the block designer, they are listed here because they cause the block to fail. The designer needs to ensure that the design will work across a wide range of process variations to ensure a robust design.

Parasitic elements are accounted for in a proper design flow. This is usually done through the use of parasitic extraction. But this is not the panacea of proper design that many believe it to be. Consider the layout of a RF device. This RF device has a corresponding model that was derived from a series of measurements of a test structure of that RF device. This layout is generated from an automated layout generator that most design kits support. This layout closely matches the model. The problem is that the measurement already accounts for some of the parasitic elements of the layout of that device, since the device that

was measured has some parasitic elements already. For instance the capacitance between the drain and source of each finger is in the measurement, as well as the overlap capacitance between the gate and source. But when the device is included in a block layout, there is more routing above and around the RF device. Currently, parasitic extraction tools cannot distinguish between what parasitic elements are already in the model, and what elements are not and need to be included in the extraction. What this can lead to is over estimation of the parasitic elements, because the parasitic elements are included in the extraction explicitly, even though they are already accounted for in the model.

This leads to the next cause of block failures: bad models. There is nothing that a designer can do about bad models. The accuracy of the device models needs to be verified before the block design begins. The most reliable way to accomplish this is to fabricate a test chip and measure each of the devices being used. This way the test chip can be tailored to be sure that the devices are measured in the same way that the devices are used in the actual chip. However, if a design can be made robust enough, the bad models can be overcome.

The last example here is incorrect test benches. This is something that experienced designers usually don't have a problem with, but new designers do. The most common error in a test bench is incompletion. This means that not everything that should be included is included. A proper test bench accounts for the input loading, the output loading, parasitic elements on the supplies and ground networks, bond wires from the package, and coupling between aggressor and victim nodes. The problem is that once all of these affects are included, the test bench is very large and takes a long time to run, and in some cases, wont even converge. The designer needs to use his or her experience to know what effects will dominate the performance, and include those in the test bench.

Errors in block design can have a large impact on the testability of a chip. For instance, if the PLL does not lock, then the receiver performance cannot be measured because the mixers cannot down-covert the signal.

SoC/Integration Issues

When the RFIC is integrated into a system-on-chip (SoC), new methods of introducing errors are found. Noise, cross talk, and verifying large digital and analog circuitry together are just a couple of the issues that SoC designs need to overcome. In addition, as the level of integration increases, the ability to test a chip becomes more difficult. This is because fewer signals are being brought out of the chip, and so fewer signals are able to be observed. Special test modes need to be implemented to allow for testing of blocks in-situ. More about SoC integration will be discussed in section 2.
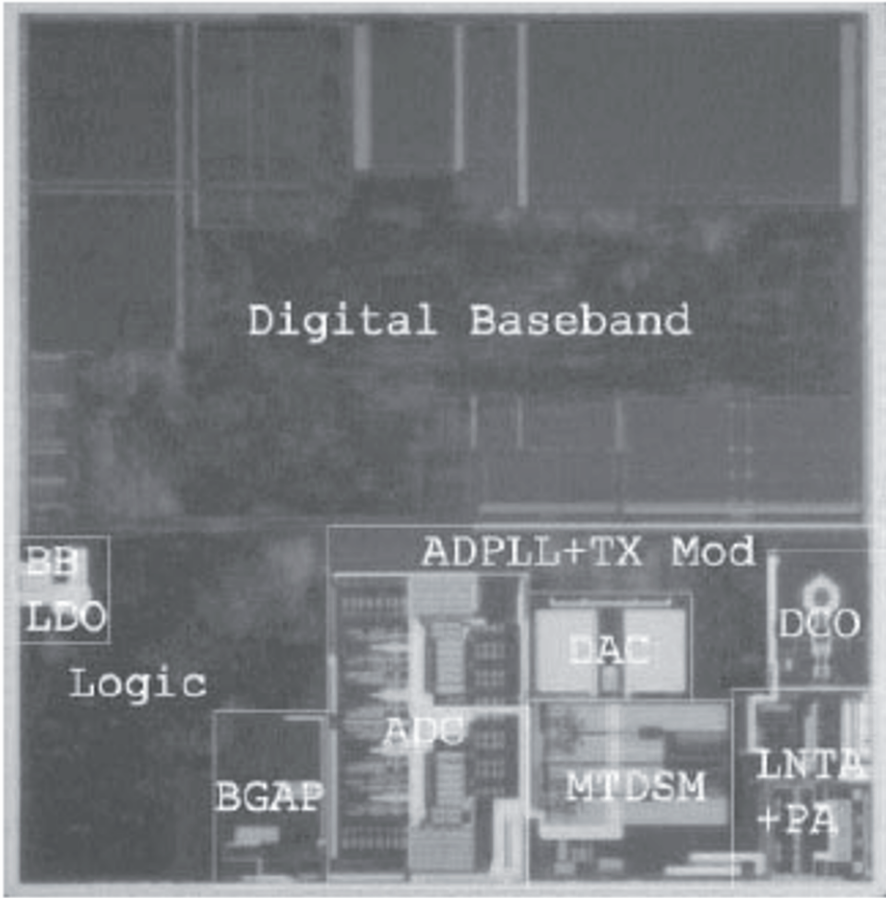
Figure 13.4.   Texas Instruments Bluetooth transceiver [3].

Yield

In a SoC design, the digital circuitry can occupy much more silicon area than analog circuitry. Digital circuits usually have a yield above 99%, while RF and analog yield is much lower. This is not as critical of an issue for a two chip design, but for a SoC it can cause a chip from going from a high margin product to a loss leader.

Consider the chip shown in figure 13.4 [3]. The digital circuitry occupies about 75% of the chip, while the analog is about 25%. Assume that the digital yield is 99% and the analog yield is 75%. If this were a two chip solution, then the yield of the digital chip would be 99% and the analog chip would be 75%, and the total amount of silicon lost to yield would be about 7%. But this is a

SoC, so the chip yield can be approximated as the lower of the analog yield and the digital yield, which means that 25% of the silicon is lost to yield problems. There is a big difference in the profitability of a product between a yield of 93% and a yield of 75%.

2. SoC Integration

This section will discuss some of the issues related to SoC integration. First the issues with RF block level integration will be discussed, followed by general integration issues. Then SoC specific issues are presented, followed by a short discussion of tapeout problems.

2.1 RF Block Integration

When integrating RF blocks, you have all the issues of analog integration, with the introduction of several new problems. One of the biggest problems with RF blocks is that the load that the block drives change during the design process. For example, an LNA drives a mixer. The mixer input impedance will change several times while the mixer is being designed. This is because the mixer input stage is being resized to meet the mixer requirements. But the LNA cannot wait for the mixer to be finished before starting design, so the LNA has to make an assumption about the mixer input impedance to begin the LNA design. At several times during the design the LNA designer will receive an update from the mixer designer as to the mixer input impedance, and adjust the test bench to account for the change. Then when the mixer is finished, the LNA must re-simulate everything to account for the mixer input impedance. It is important to simulate with the actual mixer, and not a lumped equivalent circuit for the final simulations. The lumped equivalent circuit is only valid for the frequency that the values were extracted at. So to see effects such as out of band performance, the actual mixer needs to be included to present a correct load at these far away frequencies.

Not only does the load change during the design, but the top level interconnects form a part of the load tank. These interconnects must be included in the simulation to produce accurate simulations. Usually a 3D field solver is used to model these interconnects between blocks. As an example, consider the routing between the 2.4GHz LNA and the RF mixers for the transceiver shown in figure 13.3. A close-up of the interconnect is shown in figure 13.5.

The interconnect is routed as two single paths of 5μm width, with 5μm separation. With a routing of this length, there should be a significant inductance, and because they are routed with 5μm separation there should also be large coupling, both inductive and capacitive, between them. Analyzing the interconnect in Asitic gives an inductance of 0.911nH, a resistance of 3.11Ω, and a coupling coefficient of 0.69.
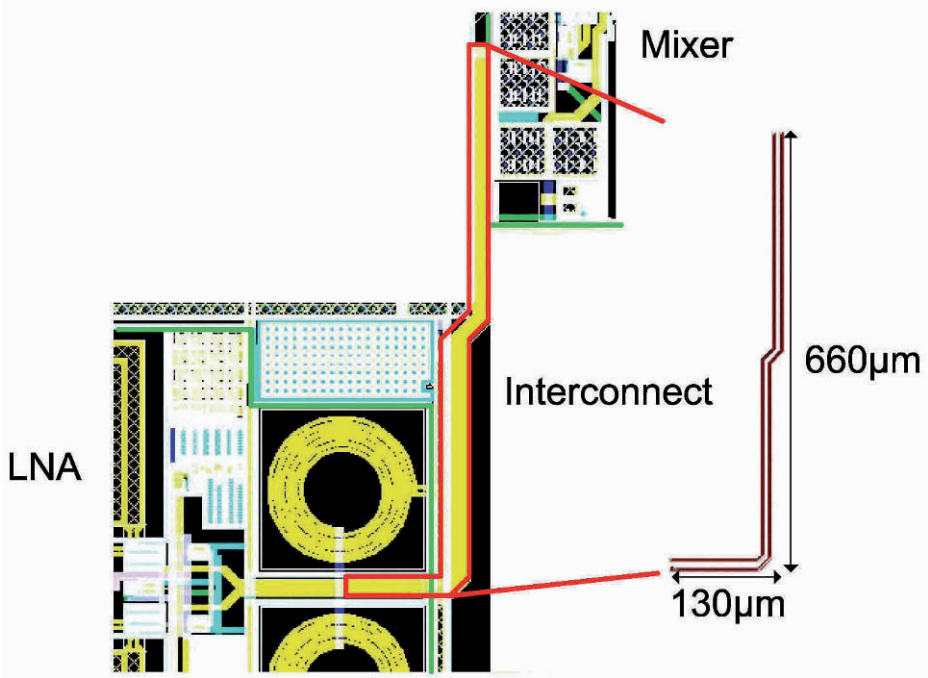
Figure 13.5.   The routing between the LNA and mixer form part of the LNA resonant tank

Notice that the horizontal routing of 130μm is included in the interconnect model? This is because it was is not included in the Asitic model of the LNA load. Even though it is only just over 100μm in length, it changes the inductance of the interconnect from 0.7 to 0.9nH.

Another interconnect effect that must be accounted for is the mutual inductive coupling between the routing and the load tank. Looking at the top inductor of the LNA, the interconnect looks like an additional quarter turn of the top load inductor, but not the lower load inductor. This will introduce some imbalance between the positive and negative paths of the differential LNA.

The 0.9nH inductance from the interconnect has a large affect on the performance of the LNA. The load inductors are provide 4.5nH of inductance at 2.4GHz, which compared to the 0.9nH of the interconnect, shows that the interconnect forms a significant part of the load tank. Without this interconnect model, the resonant frequency of the LNA would have shifted down by approximately 220MHz. Depending on the quality factor of the LNA, this could have dropped the gain by 5-10dB at 2.4GHz. In addition to this loss of gain, due to the difficulties of measuring far out of band signals in an integrated receiver, it would be very hard to find the actual frequency of resonance when testing the

receiver in the lab. Without knowing the actual resonant peak, it becomes very difficult to estimate and there for hunt down the parasitic elements that shifted the load.

2.2   General Integration Issues

One of the biggest problems with high levels of integration is noise coupling. One of the ways to mitigate noise coupling is to filter the bias network inside each of the blocks. In addition to filtering the bias network, current mirrors can be filtered also. However, this is not always possible, since a large filter on the bias network will cause the block to startup and shutdown slowly. One way to avoid this problem is to use a dual time constant filter, where at startup a small time constant is used to provide fast startup, and a large time constant is used during active operation to provide better noise isolation. Another approach to mitigate noise coupling through the bias network is to use multiple bandgaps or voltage regulators, giving each noisy block its own separate bias network.

To minimize supply coupling, each major section of the chip has its own power supply domain. The power supplies on a typical RFIC are split between receiver and transmitter, RF and baseband, PLL, input/output, and digital. This gives 7 different supply domains:

- Receiver RF supply (RXRF)

- Receiver BB supply (RXBB)

- Transmitter RF supply (TXRF)

- Transmitter BB supply (TXBB)

- Each PLL with its own supply domain (LO1, LO2, etc)

- Input/Output supply (IO)

- Digital supply (DIG)

An example of a good RF package is the quad-flat-no-lead package[4], shown in figure 13.6. It has an open die paddle, which allows for down-bonding of the grounds. This allows for as many grounds as can be fit in the pad ring of the chip. Using a large amount of ground bond wires minimizes the inductive drop in the ground network, giving a more solid ground voltage.

After the chip top level is assembled, the extra space is filled in with a capacitor filler cell. This filler cell serves two purposes: maximize the unit capacitance per μm^2, and maximize the metal density. Metal density issues are discussed in section 2.4.

Several different approaches are available to get the largest unit capacitance in a given area. The simplest approach, and also one of the most effective, is to
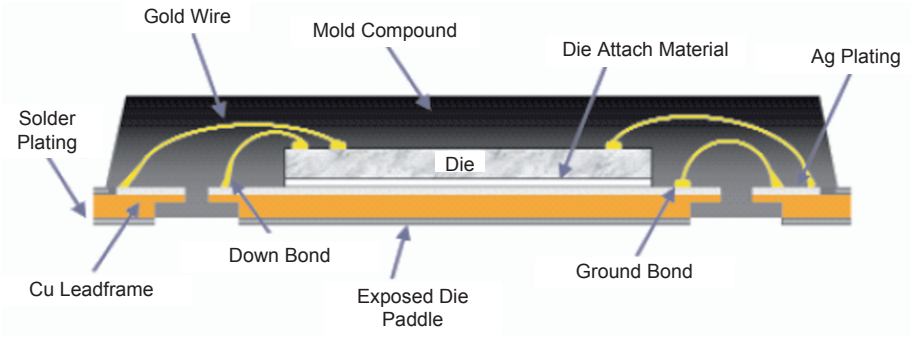
Figure 13.6. Example of a QFN package [4].
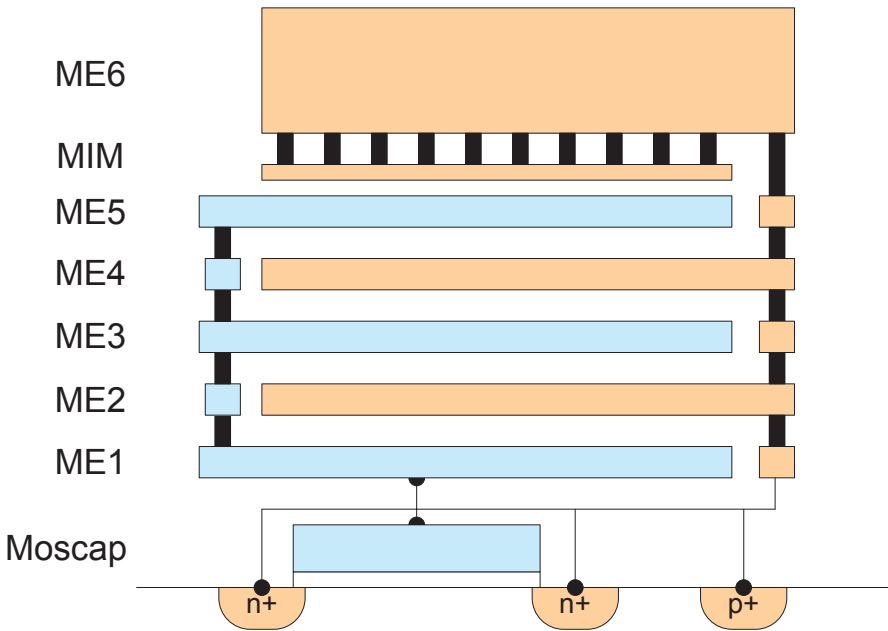
Figure 13.7. Unit cell that provides a large capacitance per μm^2 and metal density fill.

use a cell that stacks a moscap, metal overlap capacitance, and then a mim cap on top. An example is shown in figure 13.7. Using a $10\mu m$ by $10\mu m$ unit cell, the capacitance density is around 7.9fF/μm^2 on a $0.18\mu m$ process.

Even though the quality factor of the this capacitor is not very high, the structure still needs to be dampened to prevent ringing on the LRC network formed from the supply lines and decoupling. To implement the dampening, four 200Ω resistors in parallel are used to from a 50Ω resistor, which is put in series with the decoupling capacitance.

This structure allows for very high on-chip capacitance (multi-nF totals), but this may not be enough for very noisy circuits. Another solution is to use low-voltage dropout regulators (LDOs) to provide a regulated supply voltage to each independent block.

2.3   SoC Issues

There are two main issues that need to be addressed for a SoC: performance issues and functional issues. The performance issues are noise and crosstalk, specifically substrate isolation, package isolation and signal coupling, and performance verification of large mixed-signal blocks. The functional issues are mixed signal verification and pre-silicon functional validation.

Substrate Isolation

Substrate isolation is required because the substrate is shared between the RF, analog and digital circuits. The RF blocks are sensitive to coupling noise down to around 1μV, while analog blocks are sensitive to coupling noise down to around 10μV. Digital noise is couples to both analog and RF blocks, while analog noise couples to the RF.

To isolate the substrate, it is recommended to use a high resistivity substrate, on the order of 10-50Ω-sq. The high resistivity substrate provides isolation through physical separation. If a low resistivity substrate is used, then after about 20μm, physical separation no longer provides any benefits [7]. The high resistivity substrate also provides an excellent RF medium, as the high resistivity minimizes eddy currents generated from inductors in the substrate. This allows for high quality inductors to be fabricated on silicon. In addition, it gives excellent block to block isolation.

However, a high resistivity substrate is a poor digital medium. The high resistivity means that the substrate is near floating, so more substrate contacts are required to ground the bulk, and this means a larger standard cell. Also, the larger resistivity means that blocks are more sensitive to latch-up. This requires that the digital cells have a larger separation, which grows the digital size.

Another technique which can be used to provide substrate isolation is to separate the source from the bulk of the digital nmos transistor. This reduces the switching noise injected into the substrate, as the substrate is now biased by a quieter ground that the digital ground.

Yet another technique to provide substrate isolation is to specially design the digital blocks. If the digital blocks are designed so that large amount of

transistors do not all switch at the same time, the magnitude of the digital noise can be minimized. This is accomplished through the use of delay elements on the non-critical signals to distribute the clock switching time.

Mixed Signal Simulations

Simulation of large digital blocks with analog and RF is very time consuming. There is a lack of existing tools that work across different levels of abstraction. However, mixed signal simulations are required to verify block level interaction. A mix of behavioral and transistor level circuits can be used to speed up the simulation time. A high speed simulator can be used to speed up the simulation time. These simulators allow the accuracy to be set for each block depending on the targeted performance. The problem is that these simulators only work in the time domain, performing transient analysis.

To verify functionality, a large number of test vectors need to be simulated. Because of the time needed to check performance, these test vectors are not used. A behavioral model is developed for the analog and RF blocks to verify connectivity and model functionality. Then the chip-top simulations can be run with modelsim or any fast digital simulator.

2.4   Tapeout

After all of the blocks have finished layout, and the top level has been integrated together, it's almost time for tapeout. But first several additional steps are need: dummy blocking layers need to be added, p-implant blocking layers need to be added, and the die seal needs to be added.

The fabrication company will not fabricate chips that do not meet minimum metal density requirements. These rules are there to insure a uniform metallization so that the fabrications steps can be completed with the greatest accuracy. Usually these rules required minimum metal densities around 30%, depending on the layer. But with an RF chip, a large portion of the design is taken up by the inductors, which require no other metal near them to function properly. On these chips it can be very difficult to meet the minimum metal density.

To meet the metal density requirements, either the fab or the customer will need to run a dummification program to fill in the empty spaces with a dummy pattern. The program runs by looking a small window, and checking the metal density in this window. If this window does not have the minimum metal density in it, the program will insert metal in a special pattern. Then it steps over by a percentage of the window size, and repeats the operation. For digital and non-critically matched circuits, this is not a problem and does not affect the performance. But for matched circuits, the dummy pattern can create non-matched conditions do to geometry dependent etching effects. For RF circuits, the dummies can cause disturbances to the fields generated by the high frequency operation.
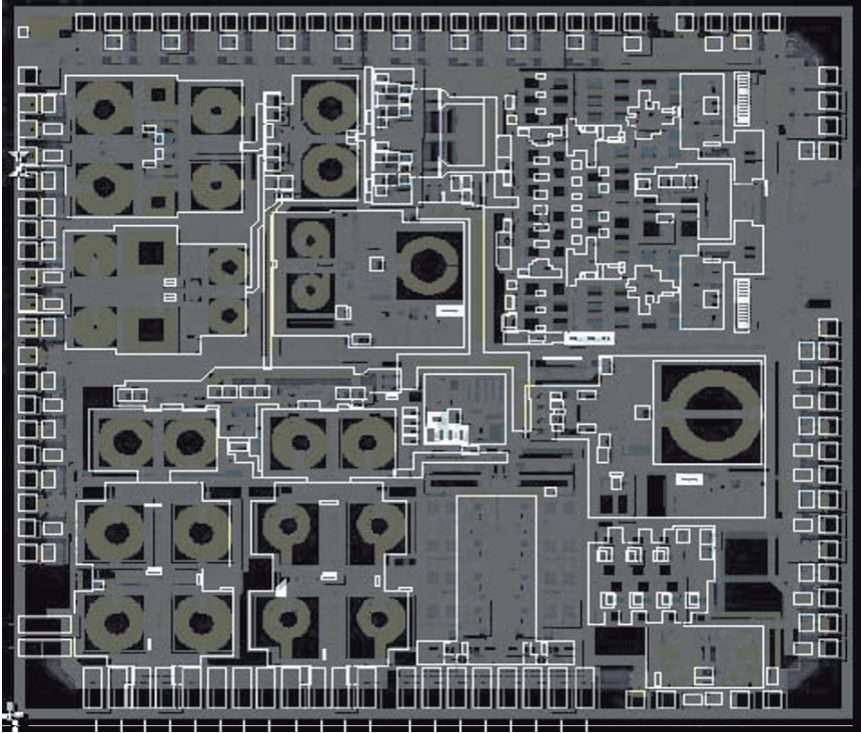
Figure 13.8. The dummy blocking layers are used to prevent dummy patterns from being added to special areas.

To avoid these problems, special layers called dummy blocking layers are used to mark areas where dummy metal should not be placed. Typically these dummy blocking layers are placed over all active circuits, inductors, mim capacitors, precision resistors and matched elements. However, adding the blocking layers to all of these circuits individually is very tedious, and can result in missed blocks. An easier way to accomplish this is to block off larger sections at a time. The drawback from this approach is that if too much of the chip is blocked off, it may be impossible to get the required metal density. Figure 13.8 shows the dummy mask for WLAN chip presented earlier in this chapter.

Similar to dummy blocking layers, a later called a p-implant blocking layer can be added to certain processes. This layer will prevent the native p-substrate from being doped to a low-resistivity substrate. The result is that areas underneath the p-implant blocking layer will be high resistivity. The problem with this is that this high resistivity substrate cannot be used to make active
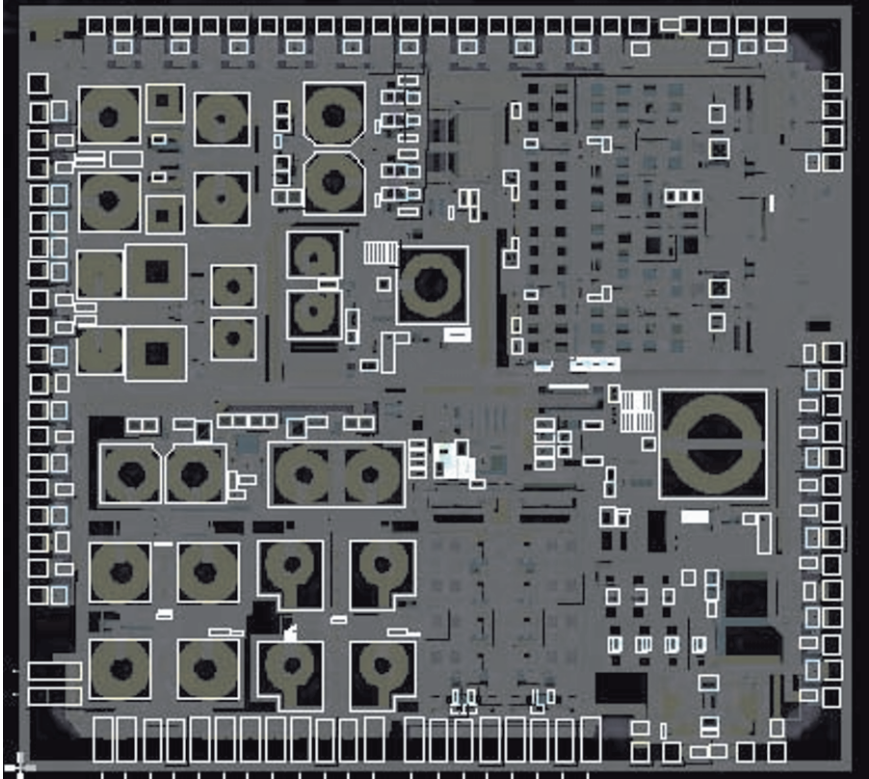
Figure 13.9.   The p-implant blocking layer is used to create high resistivity substrate regions.

devices, as no channel will form. This requires special care when placing this layer. Usually this layer is only used for certain RF applications, such as under inductors, RF mim capacitors and RF signal routings. Figure 13.9 shows the p-implant blocking layer used for the WLAN chip.

The die seal is used to prevent the substrate from cracking during the sawing of the dice. It is composed of each metal layer, one stacked on top of the other, with vias included. Special rules are used on this layout, different from the normal layout rules. A standard die seal is not used in RF design. This is because after spending so much time and effort to minimize noise coupling in the chip, a ring of metal looping around the entire perimeter would couple noise from all parts the chip around to the other parts, in effect looking like a big noise antenna. So a special die seal called a staggered die seal is employed. This die seal is actually two die seals, with breaks overlapping in each seal, similar the
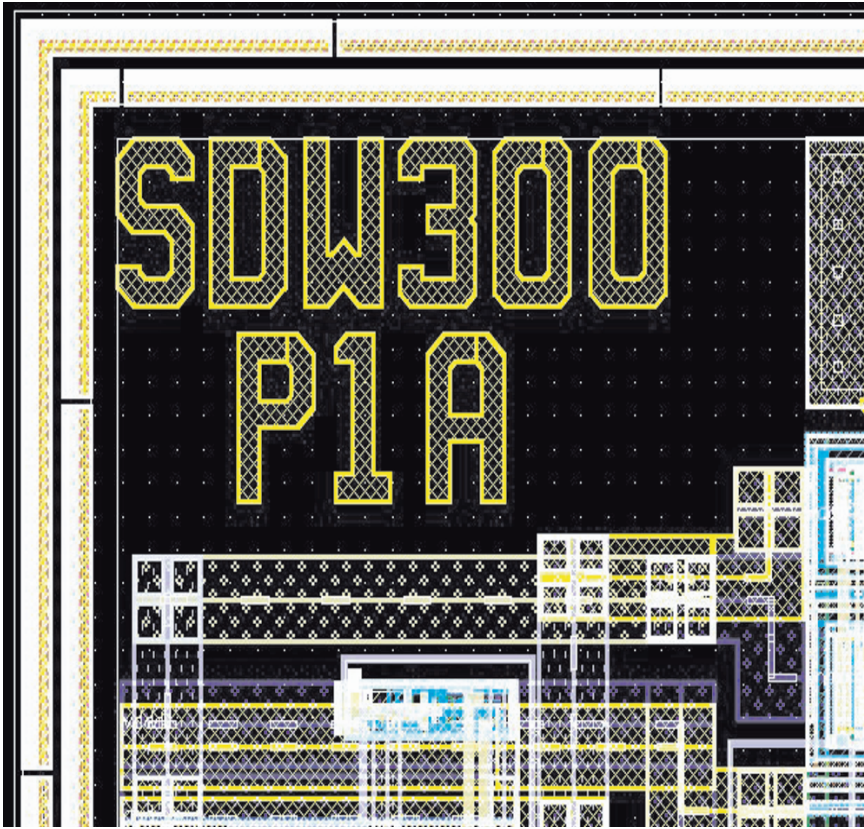
Figure 13.10.   The die seal used is a staggered die seal, designed to minimize the noise coupling in the chip.

way that bricks overlap in a brick wall. Figure 13.9 shows the principle behind the staggered die seal.

Once all of the extra layers are added, the die seal is put in place, and the chip ID has been added, the entire chip is put through a final DRC check. Once it is DRC clean, it is sent to the fab. For full mask tapeouts, usually the project lead is sent to the fab to oversee the mask creation and perform mask signoff.

As shown in figure 13.2 the cost of a mask set is very expensive. What is actually being bought are the set of glass masks that are used by the fab to generate the patterns on the wafers. There are several steps between sending the data to the fab and the actual start of wafer production.
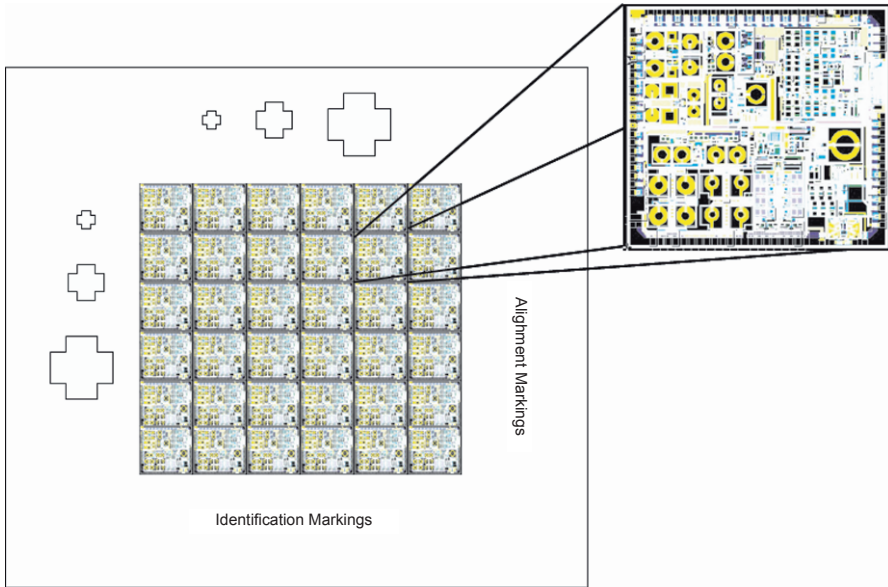
Figure 13.11.   The reticle is used to provide the stepper with an image large enough to allow for accurate steps.

Once the fab receives the data from the customer, usually in gds2 format, the fab runs its own DRC check. After they have verified that it is DRC clean, the reticle layout starts. The reticle is the layout of a series of chips next to each other to give the stepper an image large enough to use. Figure 13.11 shows a sample reticle. The reticle can be used to produce more than one version of a chip in a single full mask tapeout. For instance, if there are 6 different versions of a chip, and 2 test chips, then 50% of the reticle could be allocated to the main chip, 40% split between the 5 other high risk versions, and two spots saved for the test chips. The drawback of this approach is that the main version starts at a yield of 50%.

Between the dice on the reticle, the fab will place test structures to monitor the process. Next the fab can place the dummy patterns if the customer did not already place them. At this point the fab will run a DRC check on the whole reticle. Once the reticle passes this DRC check, it is ready to be sent to the photomask house. The photomask house will convert the data to the format that they use, usually mebes. They will run the data through their programs and produce the final mask patterns. At this point, the project lead is required to inspect the masks. The checklist that needs to be passed is very long, but the key points that are checked are:

- Check that every layer is mapped correctly, and that the tone is the correct phase

- Check that the dummy patterns are placed correctly, and on the correct layer

- Check that every layer is included

- Each die in the reticle needs to be checked, since each die used a different database

This can take from several hours to several days to check by hand, depending on the number of individual designs being produced. Once the lead designer is satisfied with the masks, the mask signoff form is signed and the data is sent down to the photomask fab to being creation of the masks.

After the wafers are completed, they are usually sent to a packaging house. The packaging house will cut the individual dice out, package them and then ship them on to the customer.

3. Self Awareness

3.1 What is Self Awareness?

Before any discussions of self awareness begin, first the question "What is self awareness" needs to be answered. Webster defines of self awareness as: "An awareness of one's own personality or individuality." Great, but how does this relate to an RFIC? Self awareness can be interpreted in two different situations for an RFIC:

1 System level self awareness

2 Block level self awareness

System level self awareness will be discusses in section 6. That leaves block level self awareness. So what does it mean that a block is self aware? This means that the block knows how well it is performing. This requires that the block is able to quantify its performance. To do this, the block needs to be able to measure its input and output characteristic. Figure 13.12 shows the idea behind a block being self aware.

This leads to two more questions:

- What parameters need to be measured?

- How does the block measure these parameters?

The answer to the first question is rather obtusely: "It depends on the block." For example, an LNA needs to measure the gain, linearity, noise and input
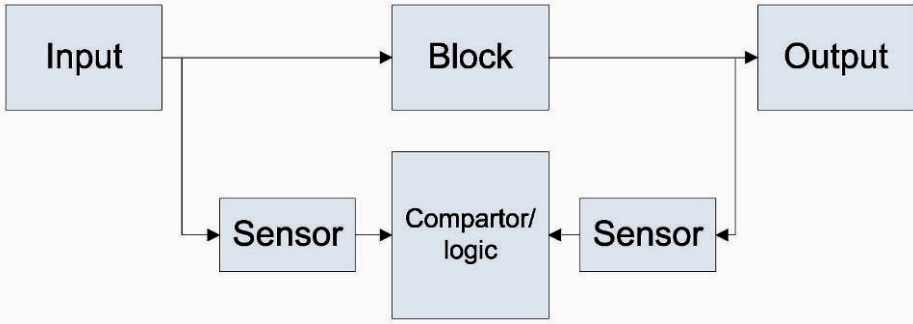
Figure 13.12. Being self aware means that the block knows how well it is performing, by measuring its input and output.

Table 13.1. Parameters to be measured for self awareness.

Block	Gain	Linearity	Noise	Input Match	I/Q	DC Offset	Filter Corner	Phase Noise	Output Power
LNA	Yes	Yes	Yes	Yes	No	No	No	No	No
Mixer	Yes	Yes	Yes	No	Yes	No	No	No	No
Baseband	Yes	Yes	Yes	No	Yes	Yes	Yes	No	No
PLL	No	No	No	No	Yes	No	No	Yes	Yes
PA	No	Yes	Yes	No	No	No	No	No	Yes

match. It doesn't care about I/Q imbalance or DC offset. Table 13.12 lists some of the more common blocks and the measurements that are needed.

The second questions answer is again "It depends on the block." For blocks line the LNA, mixer and PA, the gain and linearity are most easily measured at the input and output of the block using an RF detector. For the baseband, the parameters are most easily measured in the digital demodulator. For the PLL, the output power can be measured with the RF detector, while the phase noise can be measured in the digital baseband or using a special phase noise circuit.

3.2   Measuring the Gain

An example of a small RF detector is given in [5]. Figure 13.13 shows the schematic of the detector. When the input amplitude is low, the device is in the active region, and the output voltage is described by the drain current equation. When the input is large, the device is nonlinear. I_d becomes larger than I_{bias}, and C_2 is discharged.

The benefit of this device is that it is very low power ($2\mu A$), minimally loads the circuit (10kΩ), and has a flat frequency response. When it is used
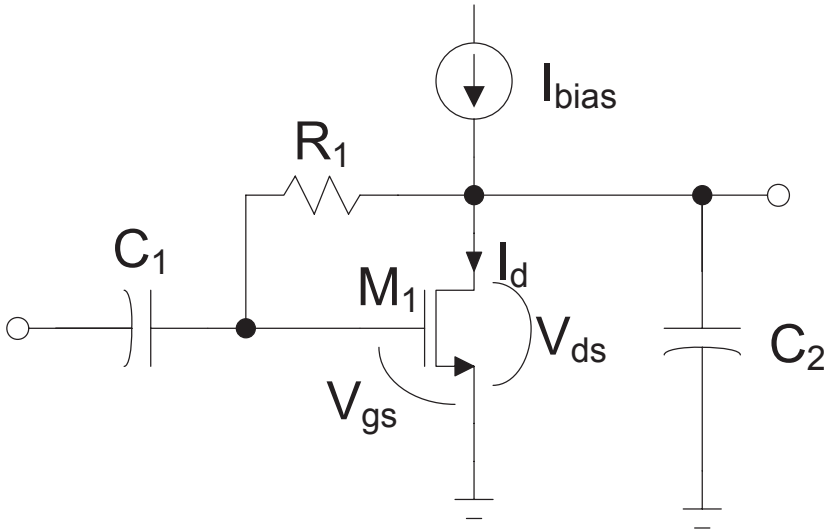
Figure 13.13. Schematic of an RF detector.
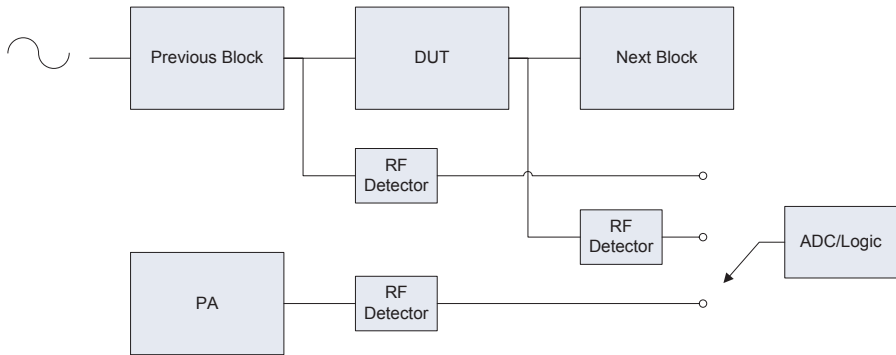
Figure 13.14. The RF detector can be calibrated against the PA to get accurate amplitude measurements.

to measure the gain of a block, the absolute accuracy of the detector is not important, so the device gives good measurements for gain. This isn't to say that the detector cannot be used to measure an absolute value, it just needs to be calibrated first. This can be done by measuring the power of a known source, such as the PA output. Figure 13.14 shows a configurations where one RF detector is calibrated first against the PA, then used to measure the input and output power of a DUT.
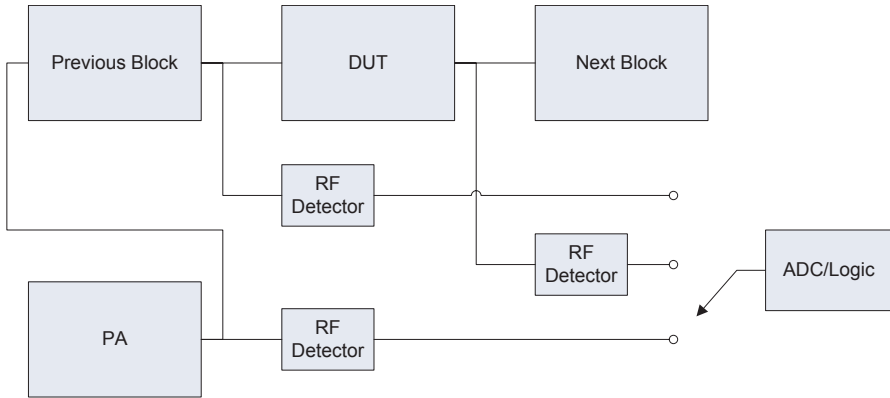
Figure 13.15.   To measure the compression point of the DUT, sweep the output power of the PA.

3.3   Measuring the Linearity

The linearity can be measured in a similar fashion as the gain. Instead of applying a fixed input power to the DUT, the PA is used to provide a swept power. Then the gain is measured and the compression point can be extracted. This is possible because the PA usually outputs much more power than the compression point of the receiver. In figure 13.15 the PA output power is swept from -30dBV to 0dBV, and the gain of the DUT is measured at each power step.

3.4   Baseband Measurements

The baseband parameters can be measured with the help of the digital baseband, which directly follows the stage. In the standard configuration shown in figure 13.16a, the I and Q imbalances can be measured. Using both ADC's the performance of an individual baseband can be measured, as shown in figure 13.16b. For instance the frequency response of the I BB is measured by using direct digital synthesis to generate a swept frequency through the I DAC, and the input and output of the I BB are then measured by the I ADC and Q ADC.

4.   Self Calibration

Now that the blocks are self aware, they each have a mechanism to track their own performance. In order to implement calibration, the cause of the impairments needs to be identified. Once they are identified, circuit techniques can be used to turn the performance of the block. This completes the calibration loop as shown in figure 13.17.
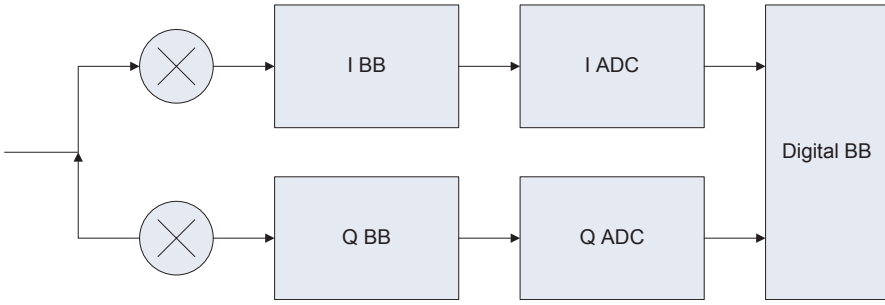
Figure 13.16a. To measure the IQ imbalance, use both receiver basebands.
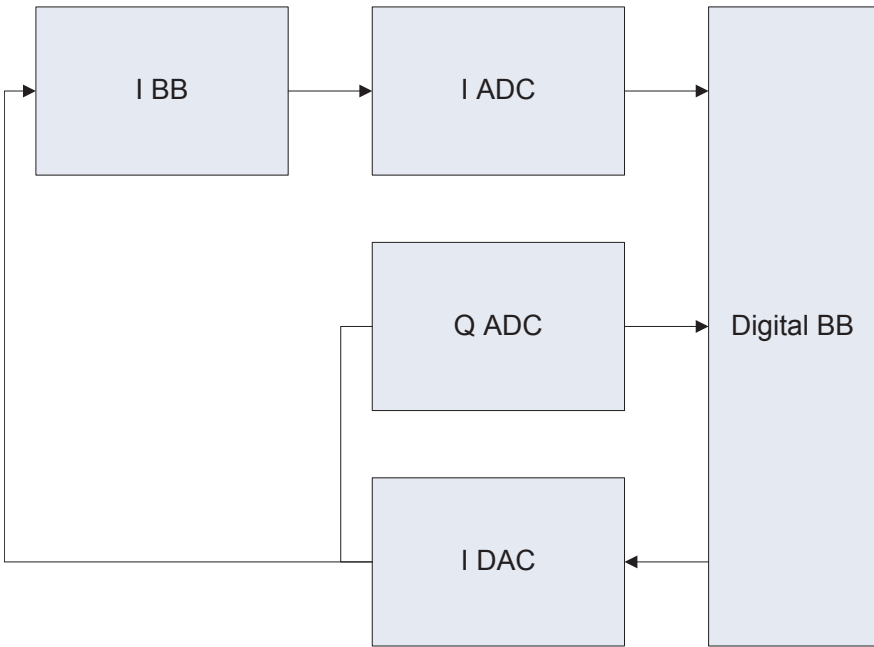
Figure 13.16b. To measure one baseband, use both ADCs and DDS from the transmitter.

The root cause of an error is usually common between similar blocks. For example, missing gain in RF blocks with tuned tanks can be caused by parasitic capacitance or inductance in those tanks. This means that different circuit techniques do not necessarily need to be created to address the same issues in different blocks. Continuing the example, to account for the parasitic elements in the tuned tanks, allow for the tank elements to be tunable.
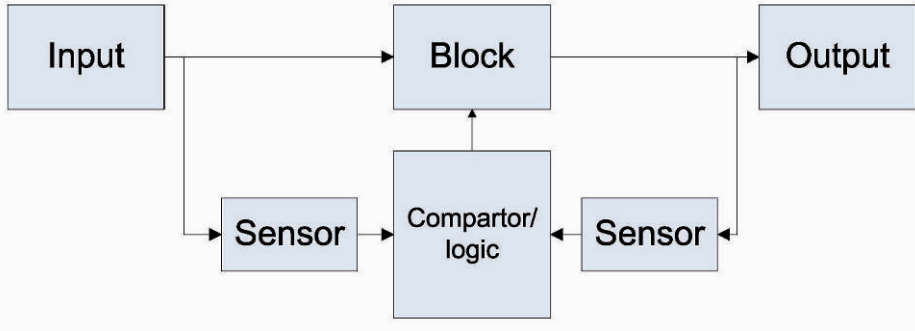
Figure 13.17. Once the cause of the impairments are found, circuit techniques can be used to close the calibration loop.

As an example, an LNA has extra capacitance added that shifts the resonant frequency of the tank. Through self awareness, the LNA knows that the peak of the gain curve has moved down by 100MHz. By tuning the load capacitance and measuring the gain again, the LNA is able to move the gain peak back to the correct frequency. Figure 13.18 shows the LNA goal and the shifted gain.

Being self aware and implementing tuning of the block performance gives bestows several important benefits. First, blocks no longer need to be over-designed to compensate for corner variations. Continuing on with the LNA example, consider the case of the slow-slow corner. To account for the lower g_m in this corner, usually the gain in the typical corner is set well above the required specification. This means that 90% of the time the LNA consumes 25% more power than necessary just in case the process shifts to the slow-slow corner. Now the LNA can increase the gain when the self awareness shows that is below the specification. Conversely, during the fast-fast corner lots, the power consumption can be reduced by lowering the gain.

Another benefit from being self aware and having self calibration is that blocks can be locally optimized based upon global parameters. If the system is operating well above the error floor, in between the noise threshold and power saturation, the power consumption in the blocks can be relaxed at the cost of noise and linearity degradation. This can drastically reduce active power consumption.

A third benefit is that multi-standard operation can be gained from a uniform architecture. By extending the length that blocks can locally tune, they can be pushed out to cover multiple standards. So if the LNA, mixer and PLL tuning range can be extended to cover 1.8GHz to 2.4GHz, the blocks can cover GSM and WLAN.
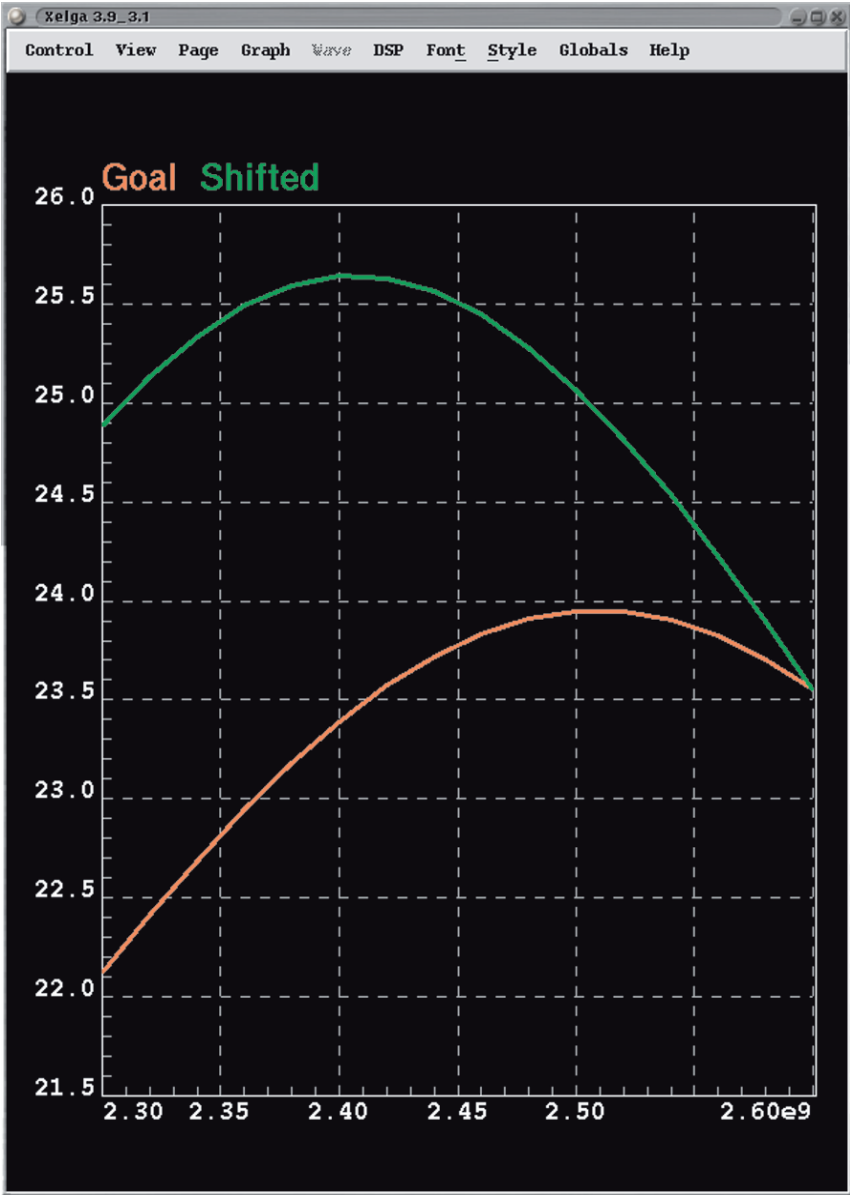
Figure 13.18. Through self awareness, the LNA knows that the gain has shifted down by 100MHz, and can retune its tanks to move it back.

5.   Self Configuration

Now that the block is self aware and can self configure, it can be designed to self configure. This means that the system can adjust the goals of the self calibration routines to change how the block is configured.

Continuing on the with the LNA example, consider this case: for the last several RX time slots, the signal has been strong and the SNR well above what is needed for the demodulator. The system tells the LNA to use channel 11, have a gain of 10dB and minimize the power consumption. During the TX-RX turnaround, the LNA adjusts the load to maximize the performance at channel 11, then reduces the power consumption down to the minimum that still produces 10dB of gain.

This feature will allow a block to adapt itself to a different set of requirements, such as different RF bands, different channel bandwidths, different data rates, etc. It is a direct extension of the self calibration that was built up before.

6.   Leveraging Self Configuration for System Parameters

Usually the most important system parameter is bit error rate (BER). The sensitivity is defined as the minimum input power that the receiver still meets the required BER. The BER must stay below the threshold in cases where blockers are present. The maximum input power is referenced to the BER in a similar manner.

To determine the sensitivity and maximum input power, link budgets are used. They are good at determining these points, but are poor for determining the error floor. These points are shown in figure 13.19.

The BER floor is better calculated through residual BER[6]. residual BER is a combination of phase noise from all the sources as well as amplifier distortion. This error floor is the region that the wireless system will operate most of the time, yet most of the design time is spent meeting the parameters for threshold and overload. If the system is operating between the threshold and overload regions, then it reasonable to assume that the BER is below the requirement for the standard. Now the system can ask the power hungry blocks to lower their power consumption at the cost of performance. This will raise the BER up close to the maximum allowable BER, while saving power.

The complete system incorporating system optimization through self configuration is shown in figure 13.20.

Two approaches can be taken to arrive at what block parameters can be changed to maximize a system parameter

1 System level analysis

2 An algorithmic approach
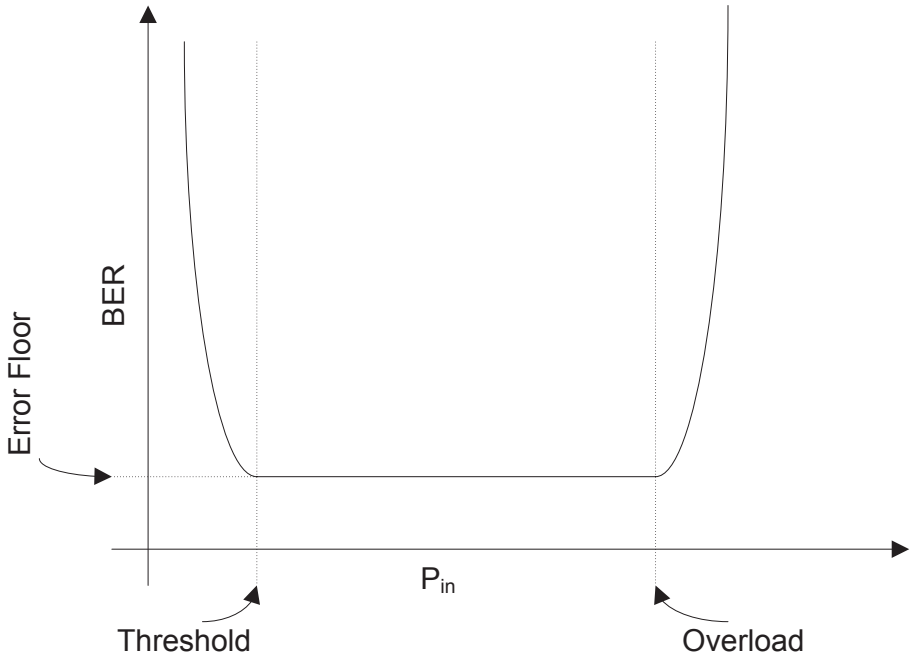
Figure 13.19. The threshold and overload points are directly related to the input sensitivity and maximum input power.
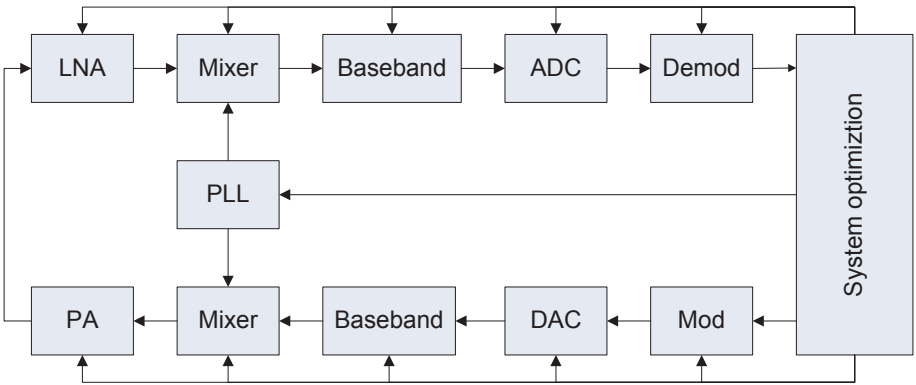
Figure 13.20. A system employing self calibration in each block to maximize performance while minimizing power consumption.

In the system level analysis approach, the performance is derived from an analytic analysis of the system blocks and modulation used. This derivation must be done for each modulation type being used, as the block non-idealities affect the BER different for each modulation type. This approach provides insight into the most critical parameters, but can be time consuming for a multi-standard SoC.

In the algorithmic approach, optimization routines are implemented in the optimization engine. These routines use existing optimization algorithms such as LMS, simulated annealing, genetic evolution, etc, to predict what the optimum parameters are. This can be performed without knowing the modulations used in advance, however it does not provide any insight into the most critical parameters.

7.   Conclusion

This chapter has covered a broad range of topics. Initially first pas silicon success was presented, and the motivation behind it discussed. Next integration issues that cause block failure in a SoC were presented. Finally, a method to provide self calibration and configuration to any block in a SoC was presented. This method provides a roadmap to not only achieving first pass silicon success, but also optimizing the performance of the radio.

References

[1] http://www.m2000.fr/products.htm

[2] Ahola, et. al, "A single-chip CMOS transceiver for 802.11a/b/g wireless LANs'," JSSC, Dec. 2004

[3] Muhammed at. al. "A discrete-time Bluetooth receiver in a 0.13μm digital CMOS process," ISSCC 2004

[4] http://www.amkor.com/Products/all_datasheets/MicroLeadFrame.pdf

[5] F. Jonsson and H. Olsson, "RF detetors for on-chip amplitue measurements," Electronic Letters, vol, 40, no. 20, Sep 2004

[6] K. K. Johnson, "Optimizing Link Budget Performance, Cost and Interchangeability by Prediciting Residual BER: Part I - Residual BER Overview and PHase Noise," *The Microwave Journal*, July 2002

[7] Gharpurey, R and Meyer, R. G., "Modeling and analysis of substrate copuling in integrated circuits," JSSC, March 1996

About the Authors

Stefan Andersson was born in Vetlanda, Sweden in 1975. He received his M.Sc. in Applied Physics and Electrical Engineering in 2000 from Linköping University in Sweden. In 2000 he joined Ericsson Microelectronics AB as designer of analog and mixed-signal circuits. Since 2001 he has been working towards his Ph.D. degree at Electronic Devices, Linköping University. He received his Licentiate Degree in 2004. During the period July to November 2004, he was working as an intern at the Intel Communication Circuit Lab, Hillsboro, Oregon, USA. His research interests are in the area of RF in silicon for wideband communication and radar applications.

Bertan Bakkaloglu joined the ASU faculty in August 2004. He received a Ph.D. in electrical and computer engineering in 1995 from Oregon State University. Prior to ASU, Dr. Bakkaloglu was with Texas Instruments where he was responsible for analog, mixed signal and RF system-on-chip development for wireless and wireline communication transceivers as a technical lead. He is a technical committee member for IEEE Radio Frequency Integrated Circuits Conference and founding chair of the IEEE Solid State Circuits Society Phoenix Chapter. Hi research interests include RF and mixed-signal IC design; data converters for wireless and wireline communication circuits and systems, integrated power management for digital communication transceivers.

Burak Çatli received the B.S. and M.S. degrees in electronic engineering from Istanbul Technical University, Istanbul, Turkey, in 1998 and 2001, respectively. He is currently working toward the Ph.D. degree in the Electrical, Computer, and Systems Engineering Department at Rensselaer Polytechnic Institute, Troy, NY. He was a graduate student and Research Assistant in the Electronic and Communication Engineering Department, Istanbul Technical University between 1998 and 2005. From 1998 to 2005, he was a Design Engineer with, the ETA-IC Design Center, Istanbul, Turkey developing RF front-end blocks and high-speed high-resolution data converter systems for industrial and military applications. His research interests are in the area of analog circuit design for integrated communication systems.

Georges G.E. Gielen received the M.Sc. and Ph.D. degrees in electrical engineering from the Katholieke Universiteit Leuven (K.U. Leuven), Leuven, Belgium, in 1986 and 1990, respectively. In 1990, he was appointed as a Postdoctoral Research Assistant and Visiting Lecturer with the Department of Electrical Engineering and Computer Science, University of California, Berkeley. In 1993, he was appointed as a tenured Research Associate with the Belgian National Fund of

Scientific Research and at the same time as an Assistant Professor with K.U. Leuven. In 2000, he was promoted to a full-time Professor with K.U. Leuven. His research interests are in the design of analog and mixed-signal integrated circuits and especially in analog and mixed-signal CAD tools and design automation. He has authored or coauthored four books and more than 250 papers in edited books, international journals, and conference proceedings. He is the Editor-in-Chief of the Elsevier Integration journal, and he is a member of the Editorial Board of the Springer International Journal on Analog Integrated Circuits and Signal Processing. He was the recipient of the 1995 Best Paper Award in the Wiley International Journal on Circuit Theory and Applications and was the 1997 Laureate of the Belgian Royal Academy on Sciences, Literature and Arts in the discipline of Engineering. He was the 2005 President of the IEEE Circuits and Systems (CAS) Society.

Mona Mostafa Hella received the B.Sc. degree with Honors from Ain-Shams University, Cairo, Egypt, in 1993, the M.Sc. in 1996, from the same university, and the Ph.D. degree, in 2001, from The Ohio-State University, Columbus, Ohio, all in Electrical Engineering. From 1993 to 1997, she was a teaching and research assistant at Ain Shams University. From 1997-2001 she was a research assistant at the Ohio-state University, working on RF circuits for wireless applications. In the summer of 1998, she was with the Helsinki University of Technology (HUT), Espoo, Finland as a visiting scholar. From June 1999-December 1999, she was with the analog group at Intel cooperation, Chandler, AZ. From 2001 to 2003, she was a senior design engineer at RF Micro Devics Inc, Billerica, MA. She is currently an assistant professor at Rensselaer Polytechnic Institute.

Ahmed Hemani is Professor at School of Information and Communications Technology, Royal Institute of Technology, Stockholm, Sweden. He got his PhD in 1993 from Royal Institute of Technology, Stockholm. His doctoral thesis on High Level Synthesis was commercialized by CADENCE as Visual Architect. Since then he has worked on many areas in System Level Design Automation including Hardware/Software Co-design, Grammar Based design methods, Hardware Synthesis from SDL, Globally Asynchronous Local Synchronous Architectures, Networks-on-Chip. He has worked in industry for ABB and National Semiconductor doing System Software Development for Embedded Systems and for Ericsson Radio Systems AB, Spirea, ACREO, NewLogic and Philips Semiconductors doing ASIC and System Design work. At Philips Semiconductors, he was Philips' representative in the Schema Working Group of the SPIRIT Consortium. His research interests include Novel System Architectures, System Level Design Automation, Reliable Systems Design and Low Power Design Methods and he has 84 publications and one patent.

Sami Hyvonen received an M.Sc. degree from Tampere University of Technology, Finland, in 2000, and a Ph.D. degree from the University of Illinois at Urbana-Champaign, in 2004 - both in electrical engineering. From 1999 to 2000 he was with Nokia Mobile Phones in Tampere, Finland, where he designed RFICs for cellular phones, and from 2000 to 2001 he was with Bell Laboratories, Holmdel, NJ, where he was doing research on RFICs for cellular base stations and WLAN transceivers. He is currently a Senior RFIC Design Engineer at Intel Corp., Hillsboro, OR.

Dr. Hyvonen has been the recipient of a Best Student Paper Award from the EOS/ESD Symposium.

Mohammed Ismail has over 20 years experience of R&D in the fields of analog, RF and mixed signal integrated circuits. He has held several positions in both industry and academia and has served as a corporate consultant to nearly 30 companies in the US, Europe and the far east.He is

The Founding Director of the Analog VLSI Lab, the Ohio State University, Columbus and of the RaMSiS Group at the Royal Institute of Technology, Stockholm. He advised the work of over 40 PhD students and of 85 Ms students. His current interest lies in research involving digitally programmable/configurable fully integrated CMOS radios with focus on low voltage/low power first-pass solutions for 3G and 4G wireless handhelds. He publishes intensively in this area and has been awarded 11 patents. He co-founded ANACAD-Egypt (now part of Mentor Graphics, Inc.) and Firstpass Technologies Inc., a developer of CMOS radio and mixed signal IPs for handheld wireless applications.

Dr. Ismail has been the recipient of several awards including the US National Science Foundation Presidential Young Investigator Award, the US Semiconductor Research Corp Inventor Recognition Awards in 1992 and 1993, and a Fulbright/Nokia fellowship Award in 1995. He is the founder of the International Journal of Analog Integrated Circuits and Signal Processing, Springer and serves as the Journal's Editor-In-Chief. He has served as Associate Editor for many IEEE Transactions, was on the Board of Governors of the IEEE Circuits and Systems Society and is the Founding Editor of "The Chip" a Column in The IEEE Circuits and Devices Magazine. He obtained his BS and MS degrees in Electronics and Communications from Cairo University, Egypt and the PhD degree in Electrical Engineering from the University of Manitoba, Canada. He is a Fellow of IEEE.

Waleed Khalil has been with Intel Corporation for over 13-years, where he held various technical leadership positions in RF and analog groups. He is currently a Sr. Staff engineer leading the frequency synthesizer design team for Intel's advanced Technology and Wireless Radio Group. Prior to that, he successfully led a group of engineers to develop Intel's first WCDMA analog front end IC. He lectured in several workshops and short courses on phase noise and is impact on wireless systems. He is a member of the paper review committee for the many IEEE Journals. He holds seven patents and has four pending. He received his BS and MS in Electrical Engineering from the University of Minnesota in 1992 and 1993, respectively. At present, he is in the process of completing his Ph.D. in the area of "Wideband frequency synthesizers and on-chip phase noise measurement techniques" at Arizona State University.

Peter Klapproth joined Philips Semiconductors in 1984. In the first ten years of his career he worked in the development of microcontrollers (8 bit as well as RISC), then at various positions in the central development of IP cores and SoC platform architectures. A constant focus point over time has been technology for on-chip communication, for which he is driving interface standardization enabling the re-use of IP cores, both in-house and external. His current focus is on advanced power management architecture development and standardization. Peter Klapproth has the degree of Dipl.-Ing. from the Technical University of Braunschweig, Germany.

Dake Liu is Professor of the chair and the Director of Computer Engineering Division, Department of Electrical Engineering at Linköping University. He received his Ph.D. (Tekn. Dr.) degree from Linköping University in Feb. 1995.

Dake Liu performs his research of Computer architecture focusing on architectures of application specific instruction set processors (ASIP) and on-chip multi-core integrations based on VLSI. His research goal is to explore different processor architectures and inter processor architectures.

Dake Liu is the co-founder of FreeHand DSP AB, (VIA tech Sweden AB, Liljeholmsstranden 5, S-117 43, Stockholm). Freehad DSP was acquired by VIA technologies in 2001. Dake Liu is the co-founder of Coresonic AB. Dake Liu was a senior ASIC designer and low power design specialist in Ericsson Microelectronics, Stockholm, Sweden (Infineon Technologies Wireless Solutions Sweden AB) since 1995 to 1998.

Dake Liu worked as a lecturer 1987-1990, and a teaching assistance 1982-1987 at Dept. of communication and control, Northern Jiaotong University, Beijing, China. His research interests in that time were EMC and reliable system design.

Sven Mattisson received his PhD in Applied Micro Electronics from Lund University in 1986. From 1987 through 1994 he was an associate professor in Applied Micro Electronics in Lund where his research was focused on circuit simulation and analog ASIC design. 1995 he joined Ericsson in Lund to work on cellular hand-set development. Presently he is with Ericsson in Lund, where he holds a position as senior expert in analog system design. Since 1996 he is also an adjunct professor at Lund University. Dr. Mattisson is one of the principal developers of the Bluetooth concept.

Delia Rodríguez de Llera González received her Telecommunications Engineering degree with a Master of Science in Electronics Engineering from the Technical University of Madrid (UPM) in 2002, a M.Sc. in Complex Adaptive Systems from Chalmers University of Technology (CTH) in Göteborg, Sweden in 2002, and a M.Sc. in System-on-Chip for Mobile Internet from the Royal Institute of Technology (KTH) in Stockholm, Sweden, in 2003. She joined CERN (European Organization for Nuclear Research), Geneva, Switzerland, in 2002 and became a CERN fellow in 2003. She pursues her Ph.D. degree at the Royal Institute of Technology from 2004 working in the field of multi-standard wireless systems design and programmable data converters. Her research interests include analog and mixed signal design, analog to digital conversion, low-voltage and low-power circuit design, and CAD tool development.

Elyse Rosenbaum received the B.S. degree (with distinction) from Cornell University in 1984, the M.S. degree from Stanford University in 1985, and the Ph.D. degree from the University of California, Berkeley in 1992. Her field of study was electrical engineering. From 1984 through 1987, she was a Member of Technical Staff at AT&T Bell Laboratories in Holmdel, NJ. She is currently a Professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign.

   Dr. Rosenbaum's present research interests include design, testing, modeling and simulation of ESD protection circuits, and design of high-speed circuits with ESD protection. She has presented tutorials on reliability physics at the International Reliability Physics Symposium, the EOS/ESD Symposium and the RFIC Symposium. She has authored or co-authored over 90 technical papers and is an editor for IEEE Transactions on Device and Materials Reliability. Dr. Rosenbaum has been the recipient of a Best Student Paper Award from the IEDM, a Technical Excellence Award from the SRC, and an NSF CAREER award.

Ana Rusu received the diploma engineer (eq. to M.Sc.) degree in electronics and telecommunications engineering from Technical University of Iasi, Romania, Ph.D. degree in electronics engineering from Technical University of Cluj-Napoca, Romania and Docent degree in Mixed-Signal Circuits from Royal Institute of Technology (KTH), Stockholm, Sweden, in 1983, 1998, and 2006 respectively. During 1983–1986 she was with Research Institute for Electronics and Telecommunications Iasi and from 1986 to 1988 she was with Territorial Computer Centre, Piatra-Neamt, Romania. Since 1988 she has been with the Technical University of Cluj-Napoca, Electronics and Telecommunications Faculty, where she was appointed as associate professor in 1999. In 2001, she joined the Royal Institute of Technology Stockholm, where she is a senior researcher in radio and the mixed-signal systems group. Her research interests include data conversion techniques for wireless communications and the design of low-voltage low-power analog, RF and mixed-signal ICs. Ana Rusu has authored or coauthored more than 50 papers in international conference proceedings and journals.

Henrik Sjöland received the M.Sc. degree in Electrical Engineering in 1994, and the Ph.D. degree in Applied Electronics in 1997, both from Lund University, Sweden, where he is currently an associate professor. He was appointed Docent in electronic circuit design in 2002. His research interests include the design and analysis of analog integrated circuits in general, and RF CMOS circuits for wireless applications in particular. He spent one year visiting the Abidi group at UCLA as a Fulbright postdoc in 1999. He is also the author of the book "Highly linear integrated wideband amplifiers - design and analysis techniques for frequencies from audio to RF", Kluwer Academic Publishers, 1999.

Christer Svensson is professor in Electronic Devices, Linköping University. He was born in Borås, Sweden in 1941 and received the M.S. and Ph.D. degrees from Chalmers University of Technology, Sweden, in 1965 and 1970 respectively. He was with Chalmers University from 1965 to 1978, where he performed research on MOS transistors, nonvolatile memories and gas sensors. He joined Linköping University 1978, and is since 1983 professor in Electronic Devices there. He initiated a new research group on integrated circuit design. Svenssons present interests are high performance and low power analog and digital CMOS circuit techniques for computing, signal processing and sensors. Svensson has published more than 170 papers in international journals and conferences and holds ten patents. He was awarded the Solid-State Circuits Council 1988-89 best paper award. He is a member of the Royal Swedish Academy of Engineering Sciences. He is a cofunder of several companies, most recently Switchcore AB and Coresonic AB.

Niklas Troedsson (SM'98, M'06) received MSc in electrical engineering in 2001 and PhD in engineering in circuit design in 2005 from Lund University. He became a member of the Circuit Design Group at the Department of Electroscience in 2000. In 2005 he launched a website for Indentro (www.indentro.com), a free monolithic inductor optimization tool. His research interest include integrated low voltage RF CMOS, PLLs, ADPLLs, Delta-Sigma Modulators, VCOs, QVCOs, phase noise analysis, and monolithic inductor modeling.

James Wilson was born in San Jose, CA. He received the B.S and M.S. degrees in electrical and computer engineering from The Ohio State University, Columbus, in 1999 and 2001, respectively, where he is currently working toward the Ph.D. degree in electrical and computer engineering. During 2000 and 2001, he was a Graduate Technical Intern with Intel Corporation, working on lowpower data converters. More recently, he was a Senior RFIC Designer with Spirea AB, Stockholm, Sweden, and Spirea Microelectronics LLC, Dublin, OH. He most recently co-founded Firstpass Technologies,Inc where he is RF Engineering Manager working on RF CMOS radios and chipsets for emerging wireless applications. His current interests lie in the area of first-pass silicon for mixed signal and RF SoC designs, CMOS LNAs and mixers, and multi-standard transceivers. He holds multiple patents.

Jens Zander received the M.S degree in Electrical Engineering and the Ph.D Degree from Linköping University, Sweden, in 1979 and 1985 respectively. 1985-90 he was a partner of SECTRA, high-tech company in telecommunication & medical systems. 1989 he was appointed professor and head of the Radio Communication Systems Laboratory at the Royal Institute of Technology, Stockholm, Sweden Since 1992 he also serves as Senior Scientific Advisor to the Swedish National Defence Research Institute (FOI). In 2000 he was appointed Scientific Director and since the beginning of 2003 he is now the Director of the Center for Wireless Systems (Wireless@KTH) at the Royal Institute of Technology.

Dr. Zander has published more than hundred scientific papers and four textbooks in Wireless communications and serves also as associate editor of several scientific journals. He serves on the board of several SMEs in the Wireless area and is a member of the Royal Swedish Academy of Engineering Sciences. His current research interests include future wireless infrastructure architectures, and in particular, related resource allocation and economic issues.

Index